WHERE MEDICAL RESPONSES FIRST GO WRONG: LOGIC-THREADED SEMANTIC TREE FOR HALLUCINATION DETECTION AND CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucinations in clinical large vision-language model responses arise when claims are unsupported by medical images, patient context, or medical knowledge, and may further cascade toxically when subsequent claims depend on earlier erroneous statements. However, existing methods largely verify clinical claims independently, overlooking both inter-claim dependencies and the multimodal contextual evidence needed to assess each claim. To address these limitations, we propose **MedFH**, a unified framework for **F**irst-**H**allucination-aware detection and correction. Specifically, MedFH explicitly models claim dependencies through logic threads over a unified semantic tree, which serves as a structural backbone for query and response sentences and claims. Building on this representation, MedFH subsequently verifies each claim against two complementary sources of evidence: grounding evidence obtained from external tools and predecessor evidence selected through logic threads. Based on the verification results, MedFH locates the first hallucinated claim and uses its direct logical successors as cascade proposals to guide revision from its source sentence onward. We also introduce **MedLHallu-Bench**, a multimodal benchmark for medical hallucination detection, first hallucination localization, and correction in radiology report generation and diagnostic reasoning. Experiments on MedLHallu-Bench show that MedFH consistently outperforms state-of-the-art baselines, and external evaluation further supports its effectiveness in correcting hallucinations in naturally generated responses. [1]

1 INTRODUCTION

Clinical large vision-language models have shown growing potential in medical VQA, radiology report generation, and diagnostic reasoning by integrating medical images with clinical language (Liu et al., 2021; Song et al., 2022). However, their deployment in clinical practice remains limited by hallucinations, which refer to statements that lack support from medical images, patient context, or medical knowledge (Zhu et al., 2025). Such hallucinations may distort clinical interpretation and undermine trust in AI-assisted workflows (Asgari et al., 2025). These risks become increasingly pronounced as responses grow longer, because later claims may depend on patient facts or earlier response claims and may therefore inherit unsupported premises. Reliable clinical use therefore requires not only detecting unsupported claims, but also accounting for the dependencies through which errors can affect subsequent claims.

Existing medical hallucination detection methods (Chen et al., 2024a; Hardy et al., 2024; Zhang et al., 2024; Liao et al., 2025a;b; Jin et al., 2026) mainly address medical VQA or radiology reports through supervised detection, uncertainty estimation, and evidence-based verification. These methods often rely on supervision tailored to specific tasks or specialized pipelines and provide limited support for tracing claim dependencies and performing localized correction. A few studies further explore medical hallucination correction through knowledge infusion or verification with counter evidence (Wu et al., 2026; Zhou et al., 2026). Beyond medical applications, prior work has explored hallucination detection at the sentence, claim, or graph level (Manakul et al., 2023; Jing et al., 2024;

[1] Code and datasets will be released.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Min et al., 2023; Chen et al., 2024b; Sawczyn et al., 2026; Fang et al., 2024) and correction through self-refinement, retrieval augmentation, or visual evidence (Madaan et al., 2023; Li & Flanigan, 2024; Yin et al., 2024). Nevertheless, existing methods still largely verify response units independently, without explicitly modeling the dependencies that make a claim contingent on patient facts or earlier response claims. Even when visual or knowledge grounding is available, it is typically used to assess the current unit itself rather than combined with the specific premises on which that unit depends. This limitation raises two key challenges for reliable hallucination detection and correction.

The **first challenge** is dependency-aware verification. A response claim may require direct grounding from the image or medical knowledge, but its reliability may also depend on specific patient facts or earlier response claims. A claim can therefore appear plausible when assessed in isolation yet become unsupported or inconsistent once the premises it relies on are considered; conversely, providing the entire query and response history may introduce irrelevant context. Effective detection therefore requires identifying the material predecessors of each claim and combining them with claim-appropriate grounding evidence. The **second challenge** arises from hallucination propagation. An early unsupported claim may be repeated, summarized, or used as a premise for later claims, leading to a cascade of downstream hallucinations. Effective correction therefore requires exploiting claim dependencies to distinguish affected downstream claims from unrelated content. Within this dependency structure, the first hallucinated claim defines the correction boundary through its source sentence, while its logical successors guide revision and reverification of the affected continuation.

To address both challenges, we propose **MedFH**, a unified framework for **F**irst-**H**allucination-aware detection and correction. MedFH organizes query and response sentences and their atomic claims into a semantic tree, which serves as a structural backbone across medical tasks while preserving their hierarchy and response order. It then constructs directed logic threads to model dependencies from relevant query facts and earlier response claims to each response claim. Incoming logic threads select relevant predecessor evidence for claim verification, which is combined with type-routed grounding evidence from external tools, while outgoing logic threads capture downstream dependencies used for correction. Once the first hallucinated claim is identified, its source sentence defines the correction boundary. MedFH preserves the preceding response, revises the continuation from this boundary, and uses the claim's direct logical successors as cascade proposals for downstream reconsideration. The revised continuation is then re-extracted and its dependency links are rebuilt for iterative reverification. To support systematic evaluation, we further introduce **MedLHallu-Bench**, a multimodal benchmark covering radiology report generation and diagnostic reasoning with sentence-level hallucination annotations, first hallucination positions, and reference corrections. Evaluations across two medical tasks and two model backbones show that MedFH consistently improves hallucination detection, first hallucination localization, and correction, while external results further support its effectiveness on naturally generated medical responses.

Our contributions are summarized as follows:

- We formulate detection and correction around the first hallucination, which defines the correction boundary for revising propagated hallucinations.
- We introduce **MedFH**, which augments a unified semantic tree with logic threads to represent dependencies among claims, verifies each claim using grounding evidence and the premises linked to it, and uses its logical successors to guide cascade correction.
- We construct **MedLHallu-Bench**, a controlled multimodal benchmark for radiology report generation and diagnostic reasoning, with annotations supporting detection, first hallucination localization, and correction.
- Experiments across two backbones show consistent improvements in detection and correction, while external evaluation shows gains on naturally generated responses.

2 METHODOLOGY

2.1 TASK FORMULATION

Given a multimodal input $X = (Q, I)$, where Q denotes a clinical query and I denotes a medical image, an LVLM generates a response $R = (S_1^R, \ldots, S_T^R)$ containing T sentences. Each response

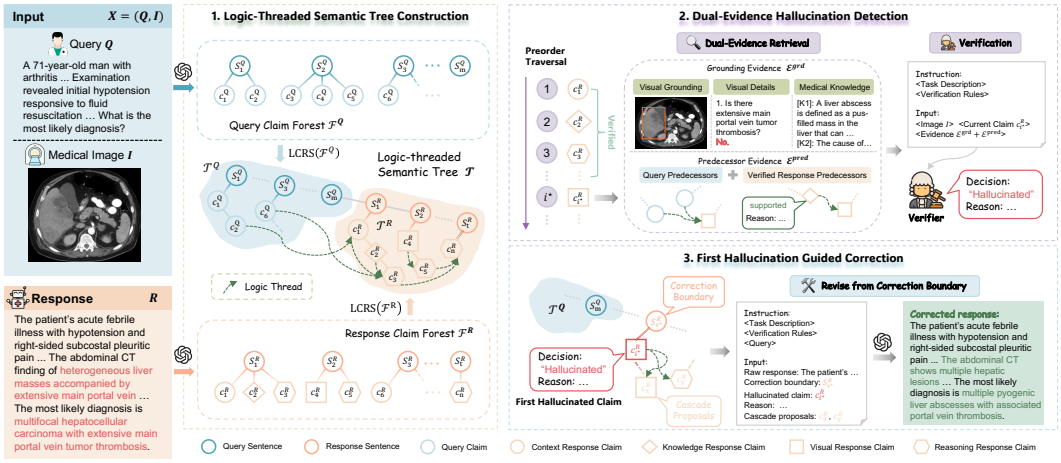


Figure 1: Overview of the proposed **MedFH** framework. Given the multimodal input $X = (Q, I)$ and an LVLM response R, MedFH first organizes the query and response forests into a unified semantic tree $\mathcal{T}$ with logic threads. It then verifies response claims in preorder using grounding and predecessor evidence. The parent sentence of the first hallucinated claim $c^R_{i^\star}$ marks the correction boundary, while its direct logical successors c^R_x, c^R_y serve as cascade proposals for revising the affected continuation.

sentence is decomposed into atomic claims, yielding $(c^R_1, \ldots, c^R_N)$ indexed by their order of occurrence in R. A hallucination detector assigns each claim a binary decision $d_i \in \{0, 1\}$, where $d_i = 1$ indicates hallucination. A sentence is labeled hallucinated if any claim originating from it is assigned $d_i = 1$.

Let $\mathrm{sent}(c^R_i)$ denote the source sentence index of c^R_i. When a hallucination is detected, the first hallucinated claim in textual order and its source sentence are defined as

$$i^\star = \min\{i \mid d_i = 1\}, \qquad \tau^\star = \mathrm{sent}(c^R_{i^\star}). \tag{1}$$

This position need not be the source of all later errors; if no hallucination is detected, the response is returned unchanged. The sentence $S^R_{\tau^\star}$ marks the correction boundary. We denote the response content before this boundary by $R_{<\tau^\star}$ and the continuation beginning at it by $R_{\geq\tau^\star}$. Correction preserves the former and revises the latter.

2.2 Framework Overview

As illustrated in Figure 1, **MedFH** instantiates this formulation in three stages. First, it decomposes the query and response into atomic claims, organizes the sentences and claims into a unified semantic tree, and adds directed logic threads from clinically relevant predecessors to response claims whose validity depends on them. Second, it traverses the response subtree in preorder and verifies each claim using type-routed grounding evidence and predecessor evidence selected through logic threads. For standalone detection, verification covers the complete response subtree and the resulting claim decisions are aggregated into sentence-level predictions. For joint detection and correction, verification stops at the first hallucinated claim. Third, MedFH revises the continuation from $\tau^\star$, using direct logical successors as cascade proposals, and rebuilds the affected response subtree and its associated logic threads for iterative reverification. The following subsections detail these stages.

2.3 Logic-Threaded Semantic Tree Construction

Claim Forest Construction. MedFH decomposes each query and response sentence into atomic medical claims. Each sentence serves as the parent of its extracted claims, forming a query claim forest $\mathcal{F}^Q$ and a response claim forest $\mathcal{F}^R$. For tasks without an explicit query, the query claim forest is empty and the query source node is retained as a structural placeholder. Query claims are treated as trusted case information, while each response claim is assigned a semantic type:

$$\kappa(c^R_i) \in \{\mathrm{visual}, \mathrm{context}, \mathrm{knowledge}, \mathrm{reasoning}\}. \tag{2}$$

The assigned type determines which external tools are used to obtain grounding evidence.

Semantic Tree Construction. The query and response claim forests are attached to dedicated source nodes under a virtual root and converted into a binary semantic tree using the standard left-child–right-sibling (LCRS) representation:

$$\mathcal{T} = \mathrm{LCRS}\big(\mathrm{Merge}(\mathcal{F}^Q, \mathcal{F}^R)\big). \tag{3}$$

The resulting tree is a deterministic semantic representation that preserves the source–sentence–claim hierarchy and textual order, forming a trusted query subtree $\mathcal{T}^Q$ and a verifiable response subtree $\mathcal{T}^R$, whose preorder traversal follows the response generation order. This ordered structure supports sequential verification and maps claim-level decisions to their source sentences.

Logic Threading. Although the binary semantic tree preserves source, hierarchy, and generation order, it does not represent clinically relevant logical relations among claims. MedFH therefore augments it with directed logic threads $\mathcal{L}$. A thread $u \rightsquigarrow v$, equivalently $(u, v) \in \mathcal{L}$, connects a query claim or an earlier response claim u to a response claim v. These threads coexist with the LCRS structure and may connect nonadjacent claims in preorder.

For each response claim c_i^R, its query and response predecessors are defined as:

$$\mathrm{Pred}_Q(c_i^R) = \{c_k^Q \mid c_k^Q \rightsquigarrow c_i^R\}, \tag{4}$$

and

$$\mathrm{Pred}_R(c_i^R) = \{c_j^R \mid j < i,\ c_j^R \rightsquigarrow c_i^R\}. \tag{5}$$

A candidate predecessor is connected to c_i^R when removing or changing it could affect the support, consistency, or interpretation of that claim. Incoming threads select relevant predecessor evidence during verification, while outgoing threads identify direct logical successors as cascade proposals. Construction details, prompts, and direct evaluations of logic threads are provided in Appendices B.1, J.1, and D.3.

2.4 Dual-Evidence Hallucination Detection

MedFH performs hallucination detection by verifying response claims sequentially in the preorder of the response subtree. For each claim, it gathers evidence from two complementary sources. Grounding evidence assesses its consistency with the medical image or established medical knowledge, while predecessor evidence evaluates its faithfulness to relevant query facts and its logical consistency with preceding claims selected through logic threads. This dual evidence verification enables MedFH to identify claims that lack grounding support or are unreliable in light of the premises on which they depend.

Type-Routed Grounding Evidence via External Tools. For each response claim c_i^R, MedFH selects external tools according to its semantic type $\kappa(c_i^R)$, and collects their outputs as grounding evidence $\mathcal{E}_i^{\mathrm{grd}}$:

- *Clinical entity grounding tool* localizes the anatomical structures and medical devices mentioned in the current claim, providing evidence of their presence and spatial correspondence in the medical image.
- *Visual detail verification tool* converts the current claim into targeted questions to examine fine-grained attributes such as laterality, severity, spatial relationships, and device position, complementing the localization evidence.
- *Medical knowledge retrieval tool* retrieves relevant passages from an external medical knowledge base to assess whether the current claim is consistent with established medical knowledge.

Visual claims use both clinical entity grounding and visual detail verification, whereas knowledge and reasoning claims use medical knowledge retrieval. Context claims invoke no external grounding tool because their validity is assessed primarily against patient-specific information. This routing strategy provides evidence appropriate to each claim type while avoiding unnecessary tool calls.

Predecessor Evidence via Logic Threads. Grounding evidence alone does not determine whether a claim is faithful to the current case or consistent with the premises it relies on. Therefore, MedFH collects two complementary forms of predecessor evidence through the logic threads of c_i^R.

Query-to-response threads select *query predecessors*, which provide the patient-specific facts and task constraints relevant to the current claim. The corresponding evidence is defined as

$$\mathcal{E}_i^Q = \mathrm{Pred}_Q(c_i^R). \tag{6}$$

Response-to-response threads select *verified response predecessors*. Because verification follows preorder, each selected predecessor has already been assigned a claim-level decision and an evidence-grounded rationale:

$$\mathcal{E}_i^R = \left\{(c_j^R, d_j, r_j) \mid c_j^R \in \mathrm{Pred}_R(c_i^R)\right\}. \tag{7}$$

Supported predecessors serve as established premises, whereas hallucinated predecessors indicate that the current claim may depend on an unreliable premise. The resulting predecessor evidence is represented as:

$$\mathcal{E}_i^{\mathrm{pred}} = \left(\mathcal{E}_i^Q, \mathcal{E}_i^R\right). \tag{8}$$

Together, $\mathcal{E}_i^Q$ anchors the current claim to patient-specific case information, while $\mathcal{E}_i^R$ evaluates its consistency with previously assessed premises. Restricting both sources to logic threads reduces interference from unrelated query and response content.

Hallucination Verification with Integrated Evidence. The verifier integrates the current claim, medical image, grounding evidence, and predecessor evidence to produce a claim-level decision and an evidence-grounded rationale:

$$(d_i, r_i) = \mathrm{Verify}\left(c_i^R, I, \mathcal{E}_i^{\mathrm{grd}}, \mathcal{E}_i^{\mathrm{pred}}\right), \qquad d_i \in \{0, 1\}, \tag{9}$$

where $d_i = 1$ denotes a hallucinated claim and r_i denotes the evidence-grounded rationale. The verified record (c_i^R, d_i, r_i) is made available to later claims for which c_i^R is a response predecessor, while claim-level decisions are aggregated into sentence-level labels according to their source sentences.

2.5 First Hallucination Guided Correction and Reverification

Correction Boundary. Once the first hallucinated claim $c_{i^\star}^R$ is identified, verification stops in the joint detection and correction setting. Its parent sentence node, corresponding to $S_{\tau^\star}^R$, marks the correction boundary. MedFH preserves the response content before this boundary and revises the continuation beginning at it.

Cascade Correction with Logical Successors. To account for hallucination propagation, MedFH collects the direct logical successors of the first hallucinated claim as cascade proposals:

$$\mathcal{P}_{i^\star}^{\mathrm{cas}} = \left\{c_j^R \mid j > i^\star,\ (c_{i^\star}^R, c_j^R) \in \mathcal{L}\right\}. \tag{10}$$

These proposals identify downstream claims whose validity may change after the first hallucination is corrected, but they are not treated as mandatory editing targets. The corrected response is generated as:

$$R^{\mathrm{corr}} = R_{<\tau^\star} \oplus \mathrm{Corr}\left(\mathbf{X}, R_{\geq\tau^\star}; c_{i^\star}^R, r_{i^\star}, \mathcal{P}_{i^\star}^{\mathrm{cas}}\right), \tag{11}$$

where $\oplus$ denotes sentence concatenation. The cascade proposals direct the correction module to reconsider downstream claims that may have inherited the error.

Iterative Reverification. After each correction, MedFH re-extracts atomic claims from the revised continuation and rebuilds its structural representation and associated logic threads. The trusted query subtree and the response content before the correction boundary remain unchanged. Verification then resumes from the correction boundary in the rebuilt subtree. If another hallucination is identified, its source sentence defines the correction boundary for the next round. This process continues until no hallucination is detected or a preset maximum number of correction rounds is reached.

Table 1: Sentence-level hallucination detection and first-hallucination localization results on MedLHallu-Bench. FLA measures exact localization of the first hallucinated sentence; **R**ecall, **P**recision, and **F1** are computed for the hallucination class, while Acc. and Mac.F1 denote overall accuracy and macro-F1. Best and second-best results are **bolded** and underlined, respectively.

Tasks	Methods	GPT-5.4-nano						Intern-S1					
		FLA ↑	R	P	F1	Acc.	Mac.F1	FLA ↑	R	P	F1	Acc.	Mac.F1
RG	VLM-as-Judge	49.69	68.06	42.32	52.19	58.98	58.13	49.54	44.99	53.42	48.84	68.52	63.05
	SelfCheckGPT	12.68	77.71	47.06	58.62	65.48	64.50	46.91	**99.62**	40.75	57.84	51.49	50.36
	FaithScore	11.88	61.32	39.12	47.77	55.89	54.79	14.04	71.17	32.01	44.16	40.78	40.56
	FActScore	42.90	58.26	38.20	46.14	55.26	53.94	36.11	32.83	40.44	36.24	62.00	54.59
	UNIHD	42.83	**95.11**	55.07	69.75	67.21	66.98	43.06	52.99	55.39	54.16	77.74	69.73
	FactSelfCheck	13.62	72.56	46.01	56.31	53.48	53.28	25.31	73.07	43.61	54.62	60.06	59.48
	MedFH	**70.17**	88.14	**72.74**	**79.70**	**85.00**	**83.90**	**65.74**	87.56	**76.46**	**81.63**	**86.84**	**85.69**
Dx	VLM-as-Judge	52.60	82.68	55.25	66.24	74.49	72.87	63.91	65.48	69.95	67.64	81.03	77.11
	SelfCheckGPT	46.17	73.50	57.94	64.80	75.86	73.22	46.52	63.25	67.44	65.28	79.66	75.45
	FaithScore	2.85	2.62	35.48	4.88	68.32	42.94	29.06	30.40	36.40	33.13	62.90	53.73
	FActScore	41.18	60.56	59.98	60.27	75.86	71.46	37.97	44.32	**93.08**	60.05	82.17	74.29
	UNIHD	54.85	**86.86**	49.77	63.28	61.68	61.61	64.59	67.59	65.70	66.63	86.25	78.99
	FactSelfCheck	28.88	65.69	34.62	45.34	52.12	51.37	33.87	68.86	42.47	52.54	62.38	60.69
	MedFH	**64.35**	77.17	**61.48**	**68.44**	**78.48**	**76.05**	**67.91**	**76.31**	79.72	**77.98**	**86.97**	**84.36**

3 Experiments and Discussion

3.1 Benchmark Construction

We construct MedLHallu-Bench which comprises two subsets. **MedLHallu-RG** contains 648 chest X-ray report cases from MIMIC-CXR (Johnson et al., 2019) and PadChest-GR (de Castro et al., 2025), and **MedLHallu-Dx** contains 561 diagnostic reasoning cases from MultiCaRe (Offidani et al., 2025), spanning X-ray, CT, MRI, and ultrasound. We standardize the source responses and inject controlled visual, factual, context, and reasoning hallucinations. When an injected hallucination is repeated, summarized, or used as a premise, the affected downstream sentences are modified consistently while unrelated content is preserved. Each case includes sentence-level hallucination labels and types, the first hallucination position, and a reference correction. Construction details, automated quality checks, artifact analysis, and statistics are provided in Appendix A.

3.2 Experimental Setup

Baselines. We compare MedFH with publicly available baselines covering representative detection and correction paradigms. Detection baselines include VLM-as-Judge and SelfCheckGPT (Manakul et al., 2023) at the sentence level, FaithScore (Jing et al., 2024), FActScore (Min et al., 2023), and UNIHD (Chen et al., 2024b) at the claim level, and FactSelfCheck (Sawczyn et al., 2026) at the graph level. Correction baselines include Self-Refine (Madaan et al., 2023), RAC (Li & Flanigan, 2024), and Woodpecker (Yin et al., 2024), representing self-refinement, retrieval augmentation, and correction guided by visual evidence, respectively. To enable comparison at the sentence level, claim, fact, and graph predictions are mapped to sentence labels, with a sentence marked as hallucinated if any constituent unit is predicted as hallucinated.

Implementation Details. We use GPT-5.4-nano and Intern-S1 (Bai et al., 2025a) as the detection and correction backbones, with the decoding temperature set to 0.0. For joint detection and correction, MedFH performs at most 3 correction rounds. MedFH uses a medically fine-tuned Qwen3-VL-4B (Bai et al., 2025b) for clinical entity grounding, Hulu-Med-7B (Jiang et al., 2025) for visual detail verification, and RRF-based retrieval over Textbooks and StatPearls collections from MedRAG (Xiong et al., 2024) for medical knowledge retrieval. DeepSeek-V4-Pro (Xu et al., 2026) is used for benchmark construction. More implementation details and inference costs are reported in Appendices B, G.

Evaluation Metrics. For detection, First Hallucination Localization Accuracy (FLA) measures the proportion of cases in which the predicted first hallucination position exactly matches the reference position. We additionally report hallucination-class precision, recall, and F1, together with overall accuracy and macro-F1 over hallucination and non-hallucination sentences. For correction,

Table 2: Hallucination correction results on MedLHallu-Bench with at most 3 correction rounds. The best and second-best results are **bolded** and underlined, respectively.

Tasks	Methods	GPT-5.4-nano			Intern-S1		
		FCA ↑	S_c ↓	AggNLG ↑	FCA ↑	S_c ↓	AggNLG ↑
RG	Self-Refine	29.17	4.12	22.25	2.16	4.90	61.07
	RAC	6.64	4.72	40.45	12.65	4.45	75.74
	Woodpecker	0.93	4.94	53.31	3.55	4.58	70.84
	MedFH	**34.62**	**3.38**	**82.08**	**37.96**	**3.46**	**80.29**
Dx	Self-Refine	33.16	3.50	24.16	26.38	3.60	52.75
	RAC	29.59	3.55	59.44	29.41	3.40	54.70
	Woodpecker	44.39	3.30	62.81	37.97	3.09	82.99
	MedFH	**57.86**	**2.32**	**82.91**	**52.23**	**2.41**	**86.07**

we report three complementary metrics. First Hallucination Correction Accuracy (FCA) measures the proportion of cases in which the annotated first hallucination is successfully resolved, using the criterion defined in Appendix B.3. The overall correction score $S_c \in [0, 5]$ assesses response-level correction quality under a fixed evaluation rubric, with lower scores indicating better correction outcomes. Following Wu et al. (2026), AggNLG measures agreement with the reference correction by averaging ROUGE-1, BERTScore, and BLEU. FCA and S_c are evaluated by Claude Sonnet 5. Full scoring and aggregation details are provided in Appendix B.3, and agreement with human ratings is reported in Appendix B.4.

3.3 Main Results and Analysis

Hallucination Detection and First Hallucination Localization. Table 1 presents the sentence-level hallucination detection and first hallucination localization results. MedFH achieves the highest FLA across all tasks and backbones, surpassing the strongest baseline by an average of 18.34 on MedLHallu-RG and 6.41 on MedLHallu-Dx. It also obtains the highest hallucination F1, overall accuracy, and macro-F1 in every setting. Exact localization is more stringent than standard detection because a false positive before the reference position causes premature correction, while a false negative at that position delays intervention. Some baselines achieve higher recall by predicting hallucinations aggressively, which reduces precision and FLA. Others obtain higher precision by flagging fewer sentences but miss more annotated hallucinations. Although MedFH does not always lead in precision or recall alone, its stronger balance yields consistently better F1 and FLA.

The gains over UNIHD suggest the value of evaluating each claim against patient context and preceding clinical statements rather than assessing it in isolation. Compared with graph-fact consistency checking, MedFH integrates claim-specific grounding with dependency-selected predecessor evidence while following response order. This comparison concerns the complete verification pipelines, rather than the structural encoding alone.

Hallucination Correction. Table 2 presents the hallucination correction results. MedFH achieves the highest FCA and AggNLG and the lowest S_c in all four settings. Relative to the strongest baseline for each metric, it improves FCA by up to 25.31 on MedLHallu-RG and 14.26 on MedLHallu-Dx, reduces S_c by up to 0.99 and 0.98, and increases AggNLG by up to 28.77 and 20.10, respectively. These results indicate that MedFH resolves the annotated first hallucination more often and achieves better response-level correction quality, while maintaining stronger agreement with the reference correction.

The baselines exhibit uneven strengths across the three correction objectives. RAC is relatively competitive in reference agreement on MedLHallu-RG with Intern-S1, whereas Woodpecker performs best among the baselines on MedLHallu-Dx. These results suggest that knowledge retrieval and visual evidence are useful, but neither baseline consistently performs well across first hallucination correction, overall correction quality, and reference agreement. In contrast, MedFH performs consistently well across all three objectives. This pattern is consistent with the benefit of locating the earliest unreliable claim, using its source sentence as the correction boundary, and reconsidering logically related downstream claims during correction.

Table 3: External correction results on the first 200 examples of each MedHEval subset. All methods use GPT-4.1-nano for correction; MedFH additionally uses GPT-5.4-nano for detection and localization. Original denotes uncorrected responses. Selection details are given in Appendix E.1.

Methods	Open-ended				Closed-ended (Acc. ↑)			
	Vis. CHAIR ↓	Vis. Recall ↑	Know. S_h ↓	Know. AggNLG ↑	MM-VisHal	CXR-VisHal	Know.	Context
Original	15.44	2.50	0.76	46.05	60.50	89.00	52.50	82.50
Self-Refine	14.97	6.46	2.00	46.84	55.50	69.50	51.50	69.00
RAC	15.47	2.50	0.76	46.08	57.50	86.50	52.00	82.50
Woodpecker	15.43	3.54	1.69	**46.92**	62.50	87.50	53.00	79.50
MedFH	**9.84**	**37.08**	**0.68**	45.34	**65.50**	**96.00**	**66.00**	**83.50**

Table 4: Component ablations on MedLHallu-Bench under the GPT-5.4-nano backbone. Dashes indicate detection outputs unchanged from MedFH by design.

Tasks	Methods	FLA ↑	Mac.F1	S_c ↓	AggNLG ↑
RG	**MedFH**	70.17	83.90	3.38	82.08
	w/o cascade proposals	-	-	3.60	81.79
	w/o first hall. corr.	-	-	3.61	76.77
	w/o logic threading	70.48	83.58	3.62	81.51
	w/o grounding evidence	40.03	69.72	3.71	79.87
	w/o verified predecessors	68.93	83.47	3.55	81.43
Dx	**MedFH**	64.35	76.05	2.32	82.91
	w/o cascade proposals	-	-	2.49	82.58
	w/o first hall. corr.	-	-	3.26	79.22
	w/o logic threading	57.22	73.11	3.08	80.06
	w/o grounding evidence	50.09	67.36	3.13	79.27
	w/o verified predecessors	55.89	69.10	3.10	80.11

Evaluation on an External Benchmark. To assess the transferability of MedFH beyond the controlled setting of MedLHallu-Bench, we evaluate it on naturally generated responses from the external MedHEval benchmark (Chang et al., 2025). MedHEval comprises open- and closed-ended tasks spanning visual, knowledge, and context hallucinations, and we follow its original evaluation protocol. As shown in Table 3, MedFH improves the original responses on seven of the eight reported metrics and achieves the best result among correction methods on each of these seven metrics. In particular, it substantially reduces visual hallucinations while improving key-finding recall, and increases accuracy across all four closed-ended subsets. The only exception is a slight decrease in knowledge AggNLG, suggesting a modest trade-off between reference similarity and hallucination removal. Overall, these results demonstrate that MedFH transfers effectively to naturally generated medical VQA responses. Adaptation details and complete results are provided in Appendix E.

3.4 Ablation Study

Table 4 systematically evaluates five key design choices. The *w/o first hall. corr.* retains all detection outputs and minimally corrects each detected error at its original location, treating the hallucinations as independent targets rather than revising the continuation from the source sentence of the first hallucinated claim. The *w/o cascade proposals* retains the same predicted first hallucination, rationale, and correction boundary, but removes its direct logical successors from the correction input. The *w/o logic threading* uses the full query context and all previously verified response premises during detection and omits cascade proposals during correction. The *w/o grounding evidence* retains the raw image but removes all external tool outputs and *w/o verified predecessors* removes response predecessor evidence while retaining query predecessors and cascade proposals.

Logic Threading and Evidence. Removing external grounding evidence causes the largest degradation in detection, underscoring the importance of claim-specific visual and medical knowledge support. Predecessor evidence complements this support, as removing verified response predecessors lowers Dx FLA from 64.35 to 55.89. However, including all predecessors rather than selecting them through logic threads yields lower Dx FLA and Mac.F1, whereas RG FLA remains similar. The benefit of predecessor selection is therefore clearer on Dx, consistent with its denser response dependencies in Appendix A.7.

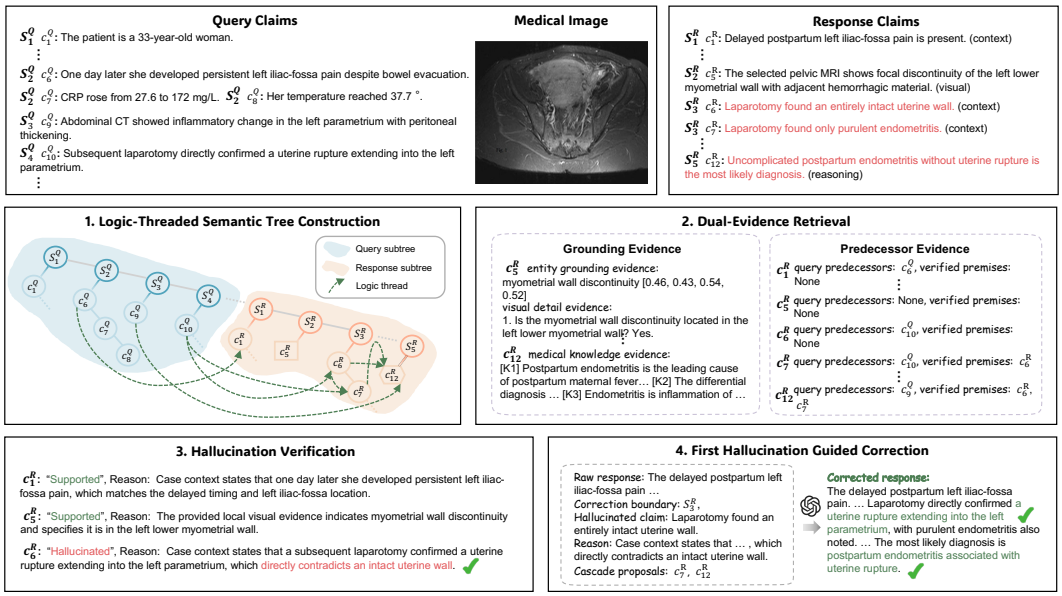


Figure 2: Qualitative case study on MedLHallu-Dx. Grounding evidence supports the visual claim c_5^R, whereas predecessor evidence reveals that c_6^R contradicts the documented uterine rupture. MedFH localizes c_6^R as the first hallucinated claim and uses its direct logical successors c_7^R and c_{12}^R as cascade proposals to correct the propagated context and reasoning errors while preserving earlier supported content.

Correction Strategy and Cascade Guidance. With detection outputs and the correction backbone fixed, first hallucination guided correction achieves a lower overall correction score than independent correction. Removing cascade proposals leaves detection unchanged but increases S_c by 0.22 on RG and 0.17 on Dx, indicating lower response-level correction quality under the same evaluation rubric. The corresponding AggNLG decreases are 0.29 and 0.33 points, respectively.

3.5 Qualitative Analysis

Figure 2 illustrates a premise-conditioned context hallucination that propagates across claim types. The visual claim c_5^R is supported by grounding evidence, whereas c_6^R, stating that laparotomy found an intact uterine wall, directly contradicts the query claim documenting a surgically confirmed uterine rupture. MedFH therefore identifies c_6^R as the first hallucinated claim. Its direct logical successors c_7^R and c_{12}^R inherit the same false premise by restricting the operative finding to endometritis and excluding rupture from the diagnosis. Using them as cascade proposals, MedFH corrects the continuation from the boundary, restores the documented rupture, updates both downstream claims, and preserves the supported history and imaging content. This case shows how incoming logic threads expose the earliest premise violation, while outgoing threads guide correction of its propagated consequences.

4 Conclusion

We presented **MedFH**, which explicitly models claim dependencies through logic threads on a unified structural backbone and verifies claims with grounding and predecessor evidence. The source sentence of the first hallucinated claim defines the correction boundary, while its logical successors guide revision of the affected continuation. We further introduced **MedLHallu-Bench**, supporting hallucination detection, localization, and correction in radiology report generation and diagnostic reasoning. Results support detection and correction on MedLHallu-Bench and correction on natural MedHEval responses. Limitations and future work are discussed in Appendix H.

REFERENCES

Vibhor Agarwal, Yiqiao Jin, Mohit Chandra, Munmun De Choudhury, Srijan Kumar, and Nishanth Sastry. Medhalu: Hallucinations in responses to healthcare queries by large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 20, pp. 31–44, 2026.

Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *NPJ digital medicine*, 8(1):274, 2025.

Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, et al. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*, 2025a.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Aofei Chang, Le Huang, Parminder Bhatia, Taha Kass-Hout, Fenglong Ma, and Cao Xiao. Medheval: Benchmarking hallucinations and mitigation strategies in medical large vision-language models. *arXiv preprint arXiv:2503.02157*, 2025.

Jiawei Chen, Dingkang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*, 2024a.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Unified hallucination detection for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3235–3252, 2024b.

Daniel Coelho de Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, et al. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *NEJM AI*, 2(7):AIdbp2401120, 2025.

Xinyue Fang, Zhen Huang, Zhiliang Tian, Minghui Fang, Ziyi Pan, Quntian Fang, Zhihua Wen, Hengyue Pan, and Dongsheng Li. Zero-resource hallucination detection for text generation via graph-based contextual knowledge triples modeling. *arXiv preprint arXiv:2409.11283*, 2024.

Romain Hardy, Sung Eun Kim, Du Hyun Ro, and Pranav Rajpurkar. Rextrust: A model for fine-grained hallucination detection in ai-generated radiology reports. *arXiv preprint arXiv:2412.15264*, 2024.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Mengyuan Jin, Zehui Liao, and Yong Xia. V-loop: Visual logical loop verification for hallucination detection in medical visual question answering. *arXiv preprint arXiv:2601.18240*, 2026.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5042–5063, 2024.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Bidur Khanal, Sandesh Pokhrel, Sanjay Bhandari, Ramesh Rana, Nikesh Shrestha, Ram B Gurung, Cristian Linte, Angus Watson, Yash R Shrestha, and Binod Bhattarai. Hallucination-aware multimodal benchmark for gastrointestinal image analysis with large vision-language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 235–245. Springer, 2025.

Changmao Li and Jeffrey Flanigan. Rac: Efficient llm factuality correction with retrieval augmentation. *arXiv preprint arXiv:2410.15667*, 2024.

Zehui Liao, Shishuai Hu, Ke Zou, Huazhu Fu, Liangli Zhen, and Yong Xia. Vision-amplified semantic entropy for hallucination detection in medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 669–679. Springer, 2025a.

Zehui Liao, Shishuai Hu, Ke Zou, Mengyuan Jin, Yanning Zhang, Huazhu Fu, Liangli Zhen, and Yong Xia. Univrse: Unified vision-conditioned response semantic entropy for hallucination detection in medical vision-language models. *arXiv preprint arXiv:2503.20504*, 2025b.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13753–13762, 2021.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

Mauro Nievas Offidani, Facundo Roffet, María Carolina González Galtier, Miguel Massiris, and Claudio Delrieux. An open-source clinical case dataset for medical image classification and multimodal ai applications. *Data*, 10(8):1–21, 2025.

Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. Medhallu: A comprehensive benchmark for detecting medical hallucinations in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2858–2873, 2025.

Albert Sawczyn, Jakub Binkowski, Denis Janiak, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. Factselfcheck: Fact-level black-box hallucination detection for llms. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 5603–5621, 2026.

Xiao Song, Xiaodan Zhang, Junzhong Ji, Ying Liu, and Pengxu Wei. Cross-modal contrastive attention model for medical report generation. In *Proceedings of the 29th international conference on computational linguistics*, pp. 2388–2397, 2022.

Jinge Wu, Zhaolong Wu, Ruizhe Li, Tong Chen, Abul Hasan, Yunsoo Kim, Jason Pui-Yin Cheung, Teng Zhang, and Honghan Wu. Error correction in radiology reports: A knowledge distillation-based multi-stage framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 39451–39459, 2026.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *Advances in Neural Information Processing Systems*, 37:140334–140365, 2024.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, 2024.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Sicheng Yang, Hangjie Yuan, Wenjun Zhang, Jinwang Wang, Yichen Qian, Weihua Chen, Fan Wang, and Lei Zhu. Clinhallu: A benchmark for diagnosing stage-wise hallucinations in medical mllm reasoning, 2026. URL https://arxiv.org/abs/2606.14697.

Jin Ye, Junlong Cheng, Jianpin Chen, Zhongying Deng, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. *arXiv preprint arXiv:2311.11969*, 2023.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

Serena Zhang, Sraavya Sambara, Oishi Banerjee, Julian Acosta, L John Fahrner, and Pranav Rajpurkar. Radflag: A black-box hallucination detection method for medical vision language models. *arXiv preprint arXiv:2411.00299*, 2024.

Nan Zhou, Ke Zou, Meng Liu, Linchao He, Jiaqi Zhu, Yi Zhang, Hu Chen, and Huazhu Fu. Hallucination detection and correction in medical vlms via counter-evidence verification. *arXiv preprint arXiv:2606.18609*, 2026.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, QingqingLong QingqingLong, Yefeng Zheng, and Xian Wu. Can we trust ai doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6748–6769, 2025.

APPENDIX

A MEDLHALLU-BENCH CONSTRUCTION AND QUALITY CONTROL

We construct **MedLHallu-Bench** to evaluate hallucination detection and correction in long-form clinical tasks. It comprises two complementary subsets. **MedLHallu-RG** contains 648 radiology report generation examples, while **MedLHallu-Dx** contains 561 diagnostic reasoning examples spanning X-ray, CT, MRI, and ultrasound. The benchmark supports sentence level hallucination detection, first hallucination localization, and continuation correction.

A.1 COMPARISON WITH EXISTING BENCHMARKS

As summarized in Table 5, representative medical hallucination benchmarks differ in task, modality, annotation granularity, and error taxonomy. Several provide response-level supervision, while others supply spans, sentences, or stage-wise reasoning traces. We map their published labels to the four broad columns only for high-level comparison; the mapping is approximate and does not imply identical annotation definitions. MedLHallu-Bench contributes the combination of multimodal long-form inputs, sentence-level first hallucination localization, four complementary error categories, and recorded downstream propagation.

Table 5: Comparison with representative annotated medical hallucination detection benchmarks (Agarwal et al., 2026; Pandit et al., 2025; Xia et al., 2024; Chen et al., 2024a; Khanal et al., 2025; Yang et al., 2026).

Benchmark	**Granularity**	**Modality**	**Hallucination Types**			
			Visual	**Factual**	**Context**	**Reasoning**
MedHalu	Span	Text		✓	✓	
MedHallu	Res.	Text		✓	✓	
CARES	Res.	Multi.	✓			
Med-HallMark	Res./Sent.	Multi.	✓		✓	
GUT-VLM	Res./Sent.	Multi.	✓			
ClinHallu	Res.	Multi.	✓	✓		✓
MedLHallu-Bench	Sent.	Multi.	✓	✓	✓	✓

A.2 SOURCE DATA

MedLHallu-RG. This subset contains chest X-ray report generation cases drawn from MIMIC-CXR (Johnson et al., 2019) and PadChest-GR (de Castro et al., 2025). Each case contains one radiograph and a source report. We preserve the clinical content of the source report while standardizing formatting that would otherwise reveal dataset-specific templates.

MedLHallu-Dx. This subset contains diagnostic-reasoning cases derived from MultiCaRe (Offidani et al., 2025). The cases span X-ray, CT, MRI, and ultrasound. Each evaluated case exposes exactly one selected image to the model.

A.3 STANDARDIZATION, HALLUCINATION INJECTION, AND REFERENCE CONSTRUCTION

The benchmark is constructed in three stages. We first standardize the source data and construct evidence grounded query and response pairs. We then inject two distinct types of hallucinations into each response under a minimal modification principle, while consistently modifying any downstream claims affected by the injected errors. Finally, we pair each hallucinated response with its corresponding response before injection to construct the reference correction. All intermediate evidence fields and textual edits are retained in the construction record.

Standardization and ground truth construction. The two benchmark tasks require different standardization procedures because their source datasets provide different forms of supervision.

For report generation, we unify the report formats from the two source datasets. Each report is standardized into four modules: *Patient Demographics*, *Examination / Modality*, *Findings*, and *Impression*. The standardizer is instructed to preserve all findings, patient information, uncertainty expressions, and diagnostic conclusions supported by the source data. It may improve grammatical continuity and remove formatting specific to an individual dataset, but it may not introduce a new medical proposition or change the certainty of an existing statement.

For diagnostic reasoning, the original MultiCaRe data are provided as captions rather than structured query and response pairs. We therefore use a large language model to extract three categories of information from each caption:

- **Visible Evidence** includes patient demographics, symptoms, examination findings, and imaging findings described for other available images. This information can be presented in the query.
- **Hidden Evidence** contains diagnostic conclusions and other answer information stated in the original caption. This information is excluded from the query and used only to construct the response.
- **Image Ground Truth** describes the findings supported by the image presented to the model. These findings are also used to construct the response rather than being directly disclosed in the query.

The construction model then generates a ground truth query and response pair from these three evidence categories. The query contains only information available to the evaluated model, whereas the response integrates the image findings, relevant medical reasoning, and the diagnostic conclusion supported by the source caption.

Hallucination injection and propagation. For each ground truth response, we randomly sample two distinct hallucination types from the following four categories:

- **Context hallucination** alters a patient specific fact provided in the query or source context, such as demographic information, symptoms, medical history, or examination findings.
- **Visual hallucination** introduces an image finding that is absent, incorrectly describes an observed finding, or changes an attribute such as location, laterality, severity, morphology, or device position.
- **Factual hallucination** introduces a medical proposition that conflicts with established clinical or biomedical knowledge, including an incorrect disease characteristic, anatomical relation, or factual association.
- **Reasoning hallucination** introduces an unsupported inference or diagnostic conclusion that does not logically follow from the available patient context, visual evidence, or preceding response claims.

Each selected type contributes one independently injected error. Injection follows a minimal modification principle: the model changes only the smallest claim or textual span needed to instantiate the target error and leaves unrelated content unchanged. In particular, content preceding the selected target remains unchanged unless it contains the second independently selected injection.

When an injected error is repeated, summarized, or used as a premise by a later claim, the affected downstream claims are modified consistently. These additional edits form a controlled propagation chain rather than separate independent errors. A downstream edit is labeled as propagated only when restoring the preceding erroneous claim to its reference form would change the support, interpretation, or validity of the downstream claim. Claims that do not depend on either injected error remain unchanged.

The construction record stores the reference claim, the corrupted claim, its sentence index, the hallucination type, and all dependent downstream edits. Independently injected errors and propagated edits are recorded in separate fields, providing an auditable mapping from each initial error to its

subsequent consequences. Because the target positions are recorded during injection, the first hallucination position does not need to be inferred retrospectively from the corrupted response.

Reference correction construction. After hallucination injection and propagation are completed, the response before injection is assigned as the reference correction for the resulting hallucinated response. Although the reference pair is assembled at the final construction stage, the reference text itself is produced before any hallucination is introduced and remains unchanged throughout injection. It is therefore never generated or rewritten by the correction evaluator. This design preserves the original evidence grounded response as the correction target and prevents evaluator specific wording preferences from defining the expected output.

Annotation categories and evidence-routing types. The benchmark annotation taxonomy and the MedFH routing taxonomy describe different properties. The former records how an injected error is instantiated (*visual*, *factual*, *context*, or *reasoning*), whereas the latter selects the evidence required to verify a response claim (*visual*, *knowledge*, *context*, or *reasoning*). A factual hallucination is typically expressed in a knowledge-routed claim, but this correspondence is not imposed as a label: routing is predicted from the claim text alone and may differ when the truth of a proposition is patient- or image-specific. Benchmark error labels are never supplied to the router or verifier.

A.4 Annotation Derivation and Reference Corrections

Sentence labels, error types, and the first hallucination position are derived directly from the recorded construction process. A sentence is labeled hallucinated if it contains an injected claim or a downstream claim modified because of that injection. The first hallucination position is the source sentence of the earliest injected claim in response order.

The standardized response before hallucination injection is retained as the reference correction. This choice separates reference creation from correction evaluation and preserves all source-supported content. The reference is used only for reference based metrics and is not provided to MedFH or the compared correction methods.

A.5 LLM-Only Source-Grounded Quality Control

Clinician adjudication is not available for the current benchmark. We therefore define a *multimodel source-grounded quality audit* for filtering ambiguous or internally inconsistent constructions. The audit checks compliance with the construction rules and agreement with the available source evidence; it is not presented as clinical expert validation.

Stratified audit set. The audit uses 200 cases, with 100 cases sampled from each subset. Sampling is stratified by hallucination type, first hallucination position, and cascade depth. Three validators—GPT 5.1, claude haiku 4.5, and Gemini 3.6 flash—inspect every case independently; GPT 5.5 adjudicates unresolved cases. Method names, intended labels, injection records, and the order of answer alternatives are hidden.

Blind validation. Each validator receives the multimodal source case, standardized response, hallucinated response, and reference correction. It answers five questions and cites the image region, patient fact, source statement, or knowledge passage supporting each decision:

1. Is the clean response supported by the available source evidence?
2. Does the hallucinated response contain an unsupported or contradicted claim?
3. What is the source sentence and semantic type of the first such claim?
4. Do the proposed downstream edits revise statements that repeat or rely on the first hallucinated claim?
5. Does the reference correction remove the target hallucination without introducing a new unsupported claim?

Table 6: Source-grounded audit outcomes (percent), together with pre-adjudication Fleiss' κ among the three validators.

Subset	Audited	Clean Supported	Hallucination Valid	First-position Agreement	Cascade Agreement	Reference Accepted	Fleiss' κ
RG	100	100.00	100.00	100.00	91.84	100.00	0.693
Dx	100	100.00	100.00	100.00	92.93	100.00	0.864

Acceptance and adjudication. A field passes when at least two of the three validators agree and their cited evidence is consistent with the source case. Cases with any unresolved field are sent to a fourth model for adjudication. A candidate is retained only if the adjudicator resolves every disputed field; otherwise it is discarded. The audit reports pass rates, percent agreement, Fleiss' κ or Krippendorff's α, and rejection reasons.

Scope of model validation. The protocol can check compliance with the construction rules, agreement on the first hallucination position, hallucination type consistency, propagation consistency, and reference fidelity. It cannot establish clinical safety, replace specialist judgment, or measure agreement with clinicians. We therefore report inter-model agreement and rejection statistics rather than clinician agreement.

Table 6 reports acceptance after adjudication and agreement before adjudication. These automated checks assess consistency with the source evidence rather than agreement with clinicians. The remaining review flags are concentrated in cascade relations, as shown in Table 7; no case is flagged for an unsupported clean response, an ambiguous first position, an invalid reference correction, or a malformed construction record. All flagged cascade cases were adjudicated before retention, and only cases whose disputes were resolved were included in the released benchmark.

Table 7: Cases sent to adjudication, grouped by the field that triggered review. Counts are review flags rather than final benchmark rejections.

Reason	RG	Dx
Unsupported clean response	0	0
Ambiguous first position	0	0
Unsupported cascade relation	10	8
Reference introduces or omits content	0	0
Other or malformed record	0	0

A.6 Injection Realism and Artifact Analysis

Table 8 shows that the injected sentences are difficult to distinguish from their clean counterparts using surface form alone. Pairwise detection accuracy is 53.00% on RG and 49.00% on Dx, with AUC values of 0.51 and 0.48, respectively, all close to chance. Tokens per sentence, lexical diversity, and punctuation frequency also remain similar before and after injection. Although the injected variants contain somewhat more negation, the evaluated blinded detector performs near chance despite this difference in negation frequency.

Table 9 separately tests whether first-hallucination localization can be explained by a positional prior. MedFH reaches 67.47% overall FLA, compared with 12.17% for always selecting the first sentence and 16.61% for a uniformly sampled position. It remains substantially stronger in the early, middle, and late bins; even in the most challenging late bin, its 41.67% FLA exceeds the random-position control by 22.77 points. These results show that MedFH outperforms the always-first and uniformly random position controls.

A.7 Benchmark Statistics

Table 10 summarizes the structural properties most relevant to MedFH. MedLHallu-Dx contains denser response-to-response dependencies, with 0.80 response threads per claim compared with 0.50 for MedLHallu-RG. This pattern is consistent with the larger first-localization gain from logic

Table 8: Injection realism and surface-artifact analysis.

Subset	Text version	Tokens/Sent.	Lexical diversity	Negation (%)	Punctuation/Sent.	Pairwise detect. acc.	AUC
RG	Clean	10.90	0.28	8.00	1.61	53.00	0.51
RG	Injected	11.56	0.27	11.00	1.56		
Dx	Clean	23.84	0.46	26.00	3.32	49.00	0.48
Dx	Injected	23.32	0.46	36.00	3.26		

Table 9: First-position controls and position-stratified FLA on MedLHallu-Bench. One RG case without a complete trace is excluded from this analysis.

Method	Overall	Early	Middle	Late
Always first	12.17	26.92	0.00	0.00
Random position	16.61	16.28	16.27	18.90
MedFH	67.47	78.02	64.03	41.67

threading observed for diagnostic reasoning in the main ablation study. MedLHallu-RG, in contrast, contains a higher proportion of cases with recorded downstream propagation (82.53% versus 42.42%), showing that successor-aware correction remains relevant to report generation even when its dependency graph is comparatively sparse. The two subsets therefore exercise complementary aspects of dependency modeling.

Table 10: Structural and propagation statistics computed from cases with complete claim and dependency traces.

Subset	Sentences per response	Response threads per claim	Cases with propagation (%)	Mean cascade depth
RG	8.43	0.50	82.53	1.00
Dx	4.83	0.80	42.42	1.00

B Implementation Details

B.1 Detailed Implementation of MedFH

Models and decoding. We use GPT-5.4-nano and Intern-S1 (Bai et al., 2025a) as the detection and correction backbones in the two main experimental settings. Deterministic model calls use temperature 0.0. In joint detection and correction, MedFH performs at most three correction rounds and stops earlier when no hallucination is detected. DeepSeek-V4-Pro (Xu et al., 2026) is used only for benchmark construction. The external evidence modules use the medically fine-tuned Qwen3-VL-4B (Bai et al., 2025b) for clinical entity grounding, Hulu-Med-7B (Jiang et al., 2025) for visual-detail verification, and reciprocal-rank-fusion retrieval over the Textbooks and StatPearls collections distributed with MedRAG (Xiong et al., 2024).

B.1.1 Logic-Threaded Semantic Tree Construction

Claim forest construction. For each case, MedFH receives the response, an optional clinical query, the attached medical image, and its modality. Response and query text are processed separately. The response extractor produces medically checkable claims. A separate typing step assigns each fixed claim exactly one primary semantic type: *visual*, *context*, *knowledge*, or *reasoning*. Each extracted claim retains its source sentence index and within-sentence order. Query text is processed by a separate extractor that returns untyped trusted facts. This separation prevents patient facts from being prematurely assigned a response-side evidence type.

In a claim forest, sentence nodes are roots and the claims extracted from each sentence are their ordered children. Claims are traversed deterministically by sentence index and then by their original

within-sentence order. The query and response claim forests are attached to dedicated source nodes under a virtual root and converted into the *left-child–right-sibling representation*. This conversion preserves source hierarchy and response order but does not itself introduce clinical dependencies.

Logic threading. Before verification, every non-knowledge response claim is processed by two independent dependency-linking calls. The query router selects relevant trusted query facts, while the response cascade linker selects only earlier response claims on which the current claim genuinely depends. The two candidate pools are never mixed within a linker call. A response edge is accepted only when removing or changing the earlier claim could require the current claim's content or assertion strength to change. Mere topical similarity, shared disease vocabulary, or membership in the same broad clinical narrative is insufficient.

For visual claims, dependency selection additionally respects the evidence source: modality or sequence, examination and time point, anatomy or view, laterality or location, target entity, and the relevant attribute, relation, or measurement slot. Unknown identifiers, self-links, and future links are removed during parsing.

B.1.2 Dual-Evidence Hallucination Detection

Type-routed grounding evidence. The semantic type determines which external evidence branch is activated. Visual claims use the attached image together with object localization and focused attribute or relation checks. Context claims invoke no external tool and are assessed against relevant patient facts and response predecessors. Knowledge claims retrieve medical passages using the claim text alone. Reasoning claims combine patient-specific predecessors with medical knowledge retrieval.

Predecessor evidence through logic threads. MedFH converts the dependency graph into two explicit evidence sets. Direct query dependencies form the case context, and direct response dependencies form the verified prefix. A response predecessor is represented by its claim text, binary verdict, and full verifier rationale.

Integrated verification. The `Verifier` selects a task- and type-specific verification interface and returns strict JSON containing the claim identifier, a binary `hallucination`/`non-hallucination` verdict, error source, error type, repair action, evidence used, and an evidence-grounded reason. Sentence-level predictions are obtained by marking a sentence as hallucinated when at least one of its claims is hallucinated.

B.1.3 First Hallucination Guided Correction and Reverification

Correction boundary and cascade proposals. When correction is enabled, the source sentence of the first hallucinated claim defines the correction boundary. The response prefix before that sentence is preserved, while the continuation beginning at the boundary is rewritten. Direct logical successors of the first hallucinated claim are supplied as cascade proposals. They indicate content that should be reconsidered, not mandatory editing targets.

Iterative reverification. After each correction, MedFH re-segments the corrected continuation, re-extracts its atomic claims, and rebuilds the affected response subtree and logic threads. The trusted query subtree and the response prefix before the correction boundary remain unchanged. Verification then resumes from the correction boundary in the rebuilt subtree. If another hallucination is detected, its source sentence becomes the boundary for the next correction round. This process stops when no hallucination is detected or the maximum of 3 correction rounds is reached.

B.2 Training Details of Qwen3-VL-4B

Medical Visual Grounding Data Construction. The clinical entity grounding model is initialized from Qwen3-VL-4B (Bai et al., 2025b) and adapted using data filtered from SA-Med2D-20M (Ye et al., 2023). We retain SA-Med2D-20M instances whose annotations can be converted into object-level localization supervision and discard records that do not yield a valid image–entity–region tuple. For each retained mask, the target region is converted to its tight bounding box and paired with the

corresponding medical entity name in an image-grounding instruction example. The number of training examples is 127,444.

LoRA Fine-tuning of Qwen3-VL-4B. The grounding model is trained with LoRA while the base Qwen3-VL-4B parameters remain frozen. The training objective is autoregressive supervision of the structured entity-localization response from the image and entity query. We use LoRA adapters with rank 32 and dropout 0.05. Training uses the AdamW optimizer with a learning rate of 0.00005 and weight decay of 0.0, and a CosineLR scheduler with a warm-up ratio of 0.03. Training runs for 5 epochs with a per-device batch size of 4, gradient accumulation over 16 steps, and an effective batch size of 256 across 4 NVIDIA RTX 4090 GPUs.

B.3 Evaluation Configuration

Evaluation units and evaluable cases. We fix the benchmark cases, response sentence boundaries, and reference labels before comparing methods. For case i, $y_{it} \in \{0, 1\}$ denotes the reference hallucination label of sentence t, and $\hat{y}_{it}^{(m)}$ denotes the corresponding prediction produced by method m. Each extracted response claim retains its source sentence identifier. A sentence is predicted as hallucinated when at least one of its associated fine-grained units is predicted as hallucinated. If a method successfully processes a case but extracts no valid fine-grained unit from a particular sentence, that sentence is assigned the non-hallucination label under the shared aggregation rule and remains in the evaluation.

Some methods may fail to produce a valid prediction for an entire case that can be aligned with the benchmark sentence identifiers, due to inference, parsing, or structured-output failures. Such cases are excluded only from the metric computation of the corresponding method rather than being removed from the benchmark globally. We denote by $\mathcal{E}_m$ the subset of benchmark cases for which method m produces a complete evaluable output.

Sentence-level detection. For each method m, we pool sentence predictions over its evaluable case set $\mathcal{E}_m$ and compute a single binary confusion matrix for each task. Let TP_m, FP_m, FN_m, and TN_m denote the corresponding counts for the hallucination class. We report:

$$P_m = TP_m/(TP_m + FP_m),$$

$$R_m = TP_m/(TP_m + FN_m),$$

$$F1_{h,m} = 2TP_m/(2TP_m + FP_m + FN_m),$$

$$Acc_m = (TP_m + TN_m)/(TP_m + FP_m + FN_m + TN_m),$$

$$F1_{\text{non-h},m} = 2TN_m/(2TN_m + FP_m + FN_m),$$

$$\text{Mac.F1}_m = (F1_{h,m} + F1_{\text{non-h},m})/2.$$

These quantities are computed from pooled counts, rather than by averaging independently calculated per-case precision, recall, or accuracy. Percent-valued metrics are multiplied by 100 only after aggregation. We provide the confusion-matrix counts so that all reported detection metrics can be recomputed.

First-hallucination localization. For each method m, let τ_i denote the reference first-hallucination sentence and $\hat{\tau}_i^{(m)}$ the predicted first-hallucination sentence for case $i \in \mathcal{E}_m$. The same special value $\varnothing$ is used when no hallucination is present or predicted. We compute

$$FLA_m = \frac{100}{N_{\text{loc},m}} \sum_{i \in \mathcal{E}_m} \mathbf{1}[\hat{\tau}_i^{(m)} = \tau_i],$$

where $N_{\text{loc},m} = |\mathcal{E}_m|$ is the number of evaluable localization cases for method m. A prediction before or after the reference sentence is incorrect. If the reference contains a hallucination but none is predicted, localization is incorrect. The controlled MedLHallu-Bench cases contain injected hallucinations; the $\emptyset$ convention also defines behavior when evaluating clean responses. Gold first-hallucination positions are used only by the evaluator and are not supplied to the correction methods.

First-hallucination correction. For each eligible corrected response, the judge assigns a score $s_i \in [0, 1]$ to the direct revised counterpart of the reference first-hallucination sentence. The judge identifies all material correction targets in that sentence, preserves its originally correct meaning, and applies the complete score anchors and mandatory caps in Appendix J.4. If the revised counterpart spans adjacent sentences, that local span is evaluated together. Later repetitions, errors in another section, and downstream inconsistencies do not change this local score; they are evaluated by the whole-response metric. We use a fixed threshold of 0.5 to convert the score into a binary correction outcome: a case is counted as successfully corrected if $s_i > 0.5$ and as unsuccessful otherwise. FCA is computed as

$$\text{FCA} = \frac{100}{N_{\text{corr}}} \sum_{i=1}^{N_{\text{corr}}} \mathbb{I}[s_i > 0.5],$$

with the same threshold applied to all methods.

Whole-response hallucination repair and preservation. The judge constructs two disjoint inventories: erroneous or unsupported content in the original response, and originally correct content supported by the available evidence. It then returns $a_i \in [0, 5]$ for residual errors or incomplete repair, including newly introduced unsupported content, and $b_{\text{loss},i} \in [0, 5]$ for collateral loss or alteration of originally correct content. Half-point ratings are permitted. Dimension A does not penalize deletion of originally correct content, and Dimension B does not penalize changes to erroneous content; the same defect is not counted in both dimensions. Under the repair-oriented rubric, lower scores are better. We compute the reported score as

$$S_c = \frac{1}{N_{\text{corr}}} \sum_{i=1}^{N_{\text{corr}}} f(a_i, b_{\text{loss},i}), \tag{12}$$

where

$$f(a, b) = 5 \left[1 - \left(1 - \frac{c(a)}{5} \right) \left(1 - \frac{c(b)}{5} \right) \right], \qquad c(x) = \min\{5, \max\{0, x\}\}. \tag{13}$$

Here, a_i and $b_{\text{loss},i}$ measure residual hallucination and correct-content loss, respectively, and N_{corr} denotes the number of successfully scored responses in the correction evaluation. For inputs in $[0, 5]$, the combination simplifies to $f(a, b) = a + b - ab/5$. The two dimensions are treated symmetrically, with no tunable weights, and $S_c \in [0, 5]$, where lower is better. The combination rule, scale, and input clipping are fixed across methods; each per-response composite score is rounded to four decimal places before averaging.

Reference agreement. AggNLG is the arithmetic mean of ROUGE-1, BERTScore, and BLEU after all three components have been expressed on the same 0–100 scale:

$$AggNLG = (ROUGE\text{-}1 + BERTScore + BLEU)/3.$$

AggNLG measures agreement with the reference correction and is not interpreted as a clinical correctness rate.

Judge configuration and invalid outputs. FCA, S_c are evaluated with claude sonnet 5, at temperature 0.0, using one valid judgment per response. The evaluator receives the fields shown in Appendix J.4; no additional image or source evidence is assumed unless it is explicitly included. The same evaluation procedure is applied to MedFH and every baseline.

MedHEval evaluation. We follow the MedHEval (Chang et al., 2025) task-specific evaluation definitions on the selected cases. For knowledge-open evaluation, its S_h is computed against the validated reference answer under the MedHEval rubric. Although both scores use a 0–5 lower-is-better scale, the MedHEval and MedLHallu-Bench rubrics are distinct.

B.4 Human Evaluation of Correction Metrics

Sample and annotator. We compared human and automated evaluations of 50 MedLHallu-RG and 50 MedLHallu-Dx responses corrected by MedFH using GPT-5.4-nano. One medically trained annotator provided FCA judgments and S_c ratings, which were compared with the corresponding Claude Sonnet 5 evaluations of the same 100 outputs. Baseline outputs were not included.

Rating procedure. For FCA, the annotator assessed whether the annotated first hallucination was corrected in its revised counterpart. For S_c, the annotator assessed residual and newly introduced errors across the complete corrected response on a 0–5 scale, following the rubric in J.4. Automated scores and rationales were hidden during human assessment.

Agreement measures. We report exact agreement and Cohen's $kappa$ for the binary FCA judgments. For S_c, we report linear-weighted $kappa$ and mean absolute error on the original 0–5 scale. Results are reported separately for MedLHallu-RG and MedLHallu-Dx and for the combined evaluation set.

Table 11: Agreement between human and automated evaluations of 50 MedLHallu-RG and 50 MedLHallu-Dx responses corrected by MedFH using GPT-5.4-nano. FCA agreement is reported as a percentage. S_c is evaluated using linear-weighted $kappa$ and MAE on the original 0–5 scale. Baseline outputs are not included.

Subset	FCA agreement (%)	FCA Cohen's κ	S_c linear-weighted κ	S_c MAE
RG	96	0.884	0.890	0.206
Dx	98	0.957	0.837	0.368
Combined	97	0.928	0.863	0.287

Results and scope. FCA agreement was 96% on MedLHallu-RG and 98% on MedLHallu-Dx, with $kappa$ values of 0.884 and 0.957, respectively. For S_c, linear-weighted $kappa$ was 0.890 on RG and 0.837 on Dx, with MAE values of 0.206 and 0.368. Across the combined set, FCA agreement reached 97% with $kappa$ of 0.928, while S_c linear-weighted $kappa$ was 0.863 and MAE was 0.287. These results show close agreement with the human ratings on the evaluated MedFH outputs.

C Baselines

C.1 Hallucination Detection Baselines

For all baselines, we follow their original settings whenever applicable and make only necessary adaptations to fit the multimodal clinical setting and the sentence-level evaluation protocol of MedFH. All methods are run on the same benchmark cases and response sentences. Evaluation uses the method-specific evaluable outputs defined in Appendix B.3. For baselines that output claim-, fact-, or graph-structured predictions, we map them to sentence-level labels using a shared conservative aggregation rule: a sentence is labeled as hallucinated if any associated fine-grained prediction is hallucinated; otherwise, it is labeled as supported. For score-based baselines, we use the decision threshold provided by the released implementation whenever available. If a method extracts no valid fine-grained unit from a sentence, that sentence is assigned the supported label under the current aggregation protocol.

VLM-as-Judge. We implement VLM-as-Judge as a direct sentence-level multimodal evaluator. For each case, the judge is given the clinical query if available, medical images, and one response sentence at a time. It is prompted to return a JSON label from supported and hallucinated, with deterministic decoding at temperature 0.0. The normalized label is directly used as the sentence-level prediction.

SelfCheckGPT. SelfCheckGPT (Manakul et al., 2023) detects hallucinations by measuring consistency between the original response and sampled responses. Following its original setting, we

generate four sampled responses for each case using the same clinical query if available and medical image, with temperature set to 1.0. Each original response sentence is compared against the sampled responses, and the resulting sentence-level score is converted into the unified granularity.

FaithScore. FaithScore (Jing et al., 2024) detects visual hallucinations by verifying fine-grained atomic facts against the input image. Following the original implementation, we first identify descriptive sub-sentences in each response, decompose them into atomic facts covering entities, relations, colors, counts, and other visual attributes, and verify each fact using the visual entailment module. A response sentence is labeled as hallucinated if any atomic fact extracted from it is unsupported by the image, and the resulting sentence-level predictions are converted into the unified granularity.

FActScore. FActScore (Min et al., 2023) decomposes generated text into atomic facts and verifies each fact against external evidence. Since MedFH targets radiology report generation and clinical diagnosis rather than open-domain biographies, we keep its decomposition-and-verification pipeline but replace the evidence source with medical reference documents. Specifically, we retrieve evidence from the MedRAG (Xiong et al., 2024) Textbooks (Jin et al., 2021) corpus using MedCPT (Jin et al., 2023). Fact-level predictions are mapped to sentence-level labels using the shared aggregation rule.

UNIHD. UNIHD (Chen et al., 2024b) is a tool-augmented hallucination detection framework that extracts verifiable claims, formulates tool-specific questions, collects evidence, and verifies claim-level hallucinations. We adapt it to our clinical setting by providing the clinical query, image evidence, and response sentence as multimodal context. For fact-oriented evidence collection, we replace the original commercial web search backend with DDGS[2] for reproducibility, while keeping the remaining pipeline unchanged. The resulting claim-level predictions are aggregated to sentence-level labels.

FactSelfCheck. FactSelfCheck (Sawczyn et al., 2026) checks whether facts in the original response are consistent with multiple sampled responses. We adopt its frequency-based KG variant, which represents response facts as subject–relation–object triples and performs graph-structured consistency checking against KGs constructed from four sampled responses generated under the same setting as SelfCheckGPT. The graph-fact-level outputs are mapped to sentence-level labels using the shared aggregation rule.

C.2 Hallucination Correction Baselines

For all baselines, we follow their released implementations whenever applicable and make only the input adaptations required for radiology report generation and diagnostic reasoning. All methods are evaluated on the same cases and hallucinated responses. Self-Refine and Woodpecker receive the medical image, the complete response, and the clinical query when available, whereas RAC operates on textual context and externally retrieved evidence without image input. No method is provided with reference responses, benchmark annotations, first hallucination positions, or other hidden evidence. Within each backbone setting, the text generation components use the same correction model with deterministic decoding at temperature 0.0, and all methods return a complete corrected response for unified evaluation.Wu et al. (2026) study radiology report proofreading through medical knowledge graph distillation and external knowledge retrieval. We do not evaluate their framework here because the knowledge resources required for its implementation, including the associated medical knowledge graph, are not publicly available, making faithful reproduction infeasible.

Self-Refine. Self-Refine (Madaan et al., 2023) iteratively improves an initial response by alternating between self-generated feedback and refinement. Following the released implementation, the correction model first reviews the current response using the available clinical query and medical image, and then rewrites it according to the generated feedback. We retain the original feedback and refinement prompt structure and few-shot examples, perform two refinement rounds, and use the response from the final round as the correction output.

[2] https://github.com/deedy5/ddgs

RAC. RAC (Li & Flanigan, 2024) performs retrieval-augmented correction by decomposing the response into atomic facts, retrieving relevant evidence, verifying each fact, and rewriting the response using the resulting facts. We retain the original breakdown, verification, statement correction, and final correction prompts. For report generation, the retrieval query is constructed from the hallucinated report, whereas the clinical query is used for diagnostic reasoning. Evidence is collected through DDGS web retrieval. RAC remains a text-only method and receives neither the medical image nor any reference or annotation field from the benchmark.

Woodpecker. Woodpecker (Yin et al., 2024) corrects visual hallucinations through five stages: key concept extraction, question formulation, visual knowledge validation, visual claim generation, and hallucination correction. Following the official implementation, we use GroundingDINO to localize extracted entities, BLIP-2 to answer visual questions using the complete image or detected regions, and QA2Claim to convert the question-answer evidence into visual claims. These claims are then supplied to the original refinement module, together with the available clinical context, to generate the complete corrected response.

D Intrinsic Evaluation of Intermediate Components

D.1 Claim Extraction Quality

Claim extraction is audited on the same stratified set used for intermediate-component evaluation, while the judges see only the source sentence and its extracted claims. Three model families independently assess faithfulness, proposition coverage, atomicity, and source-sentence alignment. Faithfulness requires every extracted proposition to be licensed by the source sentence, coverage requires the clinically relevant propositions to be retained, atomicity requires each unit to support a single verification decision, and alignment checks that every claim remains attached to its original sentence. Majority-vote labels are used only for this intrinsic evaluation and are never supplied to MedFH.

As summarized in Table 12, source-sentence alignment reaches 100.00% on both subsets. MedLHallu-RG also achieves 93.91% faithfulness, 84.94% coverage, and 94.35% atomicity. Diagnostic reasoning is more compositionally demanding, but its extracted claims still retain 83.33% faithfulness and 79.38% coverage with perfect source alignment. The lower atomicity on MedLHallu-Dx reflects the greater frequency of sentences that jointly express observations, background knowledge, and patient-specific conclusions, and motivates the task-aware decomposition used by MedFH.

Table 12: Source-grounded claim-extraction quality on the 200 audited cases; all quality values are percentages.

Subset	Sentences	Faithful	Coverage	Atomic	Alignment
RG	870	93.91	84.94	94.35	100.00
Dx	479	83.33	79.38	72.52	100.00

D.2 Semantic Type Routing Quality

The same judges assign each audited response claim to visual, context, knowledge, or reasoning according to the evidence required for verification. Their consensus label is compared with the routing output. A mixed sentence is not forced into a single type before decomposition; instead, each atomic claim receives one primary type.

The confusion matrix in Table 13 is strongly concentrated on the diagonal. Visual and context claims are particularly well separated, while the remaining off-diagonal cases mainly involve reasoning claims and claims whose verification also requires either image evidence or medical knowledge. This pattern is consistent with the intrinsically soft boundary between recalling a medical fact and applying that fact to a particular patient, rather than indicating systematic routing to an unrelated evidence source.

The per-type results in Table 14 further show that visual routing achieves 99.31% precision and the highest F1 score of 88.95%. Context routing also remains strong, with 89.06% recall and an F1 score of 80.06%. Knowledge and reasoning claims are more challenging because applying general medical knowledge to a patient-specific case creates a naturally soft semantic boundary. Nevertheless, reasoning recall reaches 79.63%, indicating that most inferential claims are routed to the more comprehensive evidence pathway rather than to a narrower directly grounded pathway.

Table 13: Four-way semantic-routing confusion matrix; rows are three-model consensus labels and columns are MedFH predictions.

Gold	Visual	Context	Knowledge	Reasoning
Visual	1002	49	44	149
Context	7	285	2	26
Knowledge	0	8	80	42
Reasoning	0	50	5	215

Table 14: Per-type semantic-routing precision, recall, and F1 against three-model consensus labels; values are percentages.

Type	Support	Precision	Recall	F1
Visual	1244	99.31	80.55	88.95
Context	320	72.70	89.06	80.06
Knowledge	130	61.07	61.54	61.30
Reasoning	270	49.77	79.63	61.25

D.3 Logic Thread Quality

Logic-thread quality is evaluated in two complementary ways. First, three source-grounded judges independently label whether each predicted predecessor is necessary for assessing the current claim and identify missed predecessors from the allowed candidate set. Second, benchmark propagation records provide objective provenance for edges from the first injected error to modified downstream claims. The latter supports cascade-successor recall and path coverage without requiring clinical adjudication.

Table 15 shows that response-to-response dependencies are recovered with balanced precision and recall (83.72% and 82.39%), yielding an F1 score of 83.05%. Query-to-response links achieve 97.36% recall, indicating that the thread constructor rarely omits patient facts that may be needed for verification. Their lower precision reflects the deliberately recall-oriented query-side policy, which retains potentially relevant case evidence before the verifier determines whether it is decisive. Consequently, the combined graph preserves 91.33% of required predecessors while maintaining an overall F1 score of 71.19%.

The objective propagation analysis in Table 16 confirms that the predicted threads capture benchmark-defined cascades across both tasks. Successor recall exceeds 90% for both MedLHallu-RG and MedLHallu-Dx, and path coverage reaches 90.31% and 95.74%, respectively. These results show that logic threads recover most recorded downstream edits in the controlled benchmark.

Table 15: Precision, recall, and F1 of predicted query-to-response and response-to-response logic-thread edges; values are percentages.

Edge set	Precision	Recall	F1
Q→R	49.73	97.36	65.83
R→R	83.72	82.39	83.05
All predicted edges	58.33	91.33	71.19

Table 16: Coverage of benchmark-recorded downstream propagation by predicted logic threads; successor precision, successor recall, and path coverage are reported as percentages.

Subset	Cases	Successor Precision	Successor Recall	Path Coverage
MedLHallu-RG	534	93.54	90.20	90.31
MedLHallu-Dx	238	84.34	91.86	95.74

E External Benchmark Adaptation and Fairness

E.1 Adapting MedHEval to MedFH

MedHEval (Chang et al., 2025) evaluates naturally generated medical responses under visual-misinterpretation, knowledge-deficiency, and context-misalignment settings. Its closed-ended subsets contain binary and multiple-choice VQA, whereas its open-ended subsets contain chest-X-ray reports and explanatory knowledge answers. Unlike MedLHallu-Bench, MedHEval does not annotate the sentence-level position of the first hallucination or the downstream claims affected by it. We therefore use it to evaluate complete correction systems on natural model outputs rather than to evaluate first-hallucination localization against an external gold boundary. The adapter changes only the task interface: it reuses the main MedFH pipeline for claim traversal, semantic typing, evidence routing, dependency linking, first-error detection, continuation revision, and post-repair verification.

Task-aware claim extraction. The principal adaptation is the semantic interface between each MedHEval question–answer pair and the claim representation used by MedFH. The extractor receives the complete question, the answer, and all answer choices when choices are available. It does not receive the dataset name, the benchmark hallucination category, or a presumed claim type. This prevents the benchmark taxonomy from being treated as a verification label: for example, a question in the knowledge subset may still contain a directly observable visual claim or a patient-specific inference.

For closed-ended VQA, the selected answer is expanded into a self-contained claim that preserves the question polarity, declared modality and anatomy, and the complete answer space. Binary answers are represented together with the proposition to which "Yes" or "No" refers. Multiple-choice answers retain both the selected option and the competing choices, so verification compares candidates within the original answer space and revision can return only a legal option or binary label. Query claims separately encode patient history, findings from other examinations, and other facts stated in the question.

For open-ended knowledge VQA, the extractor separates three semantic roles that frequently occur in the same answer: general definitions or typical manifestations are represented as knowledge claims; statements about the current image are represented as visual claims; and conclusions applying general knowledge to the current patient are represented as reasoning claims. Dependency links connect an image premise or trusted query fact to the corresponding patient-specific conclusion. Consequently, a correct medical definition can be retained even when its application to the displayed image is unsupported.

For open-ended visual VQA, which is formulated as chest-X-ray report generation, claims are extracted in report order and retain their source sentence indices. Independently verifiable findings, negated findings, and device statements are separated when they have different truth conditions, while laterality, anatomical location, measurement, severity, and other attributes remain attached to the finding they modify. Diagnostic interpretations are represented as reasoning claims. This decomposition allows a localized finding to be corrected without rewriting unrelated observations; when it supports an impression-level conclusion, the dependency graph also makes that conclusion eligible for synchronized revision.

Evidence provenance and structured revision. The adapted verifier receives the original question and answer, the current claim, and only the evidence relevant to that claim. Evidence is explicitly separated into question context, the source image, localization candidates, targeted visual observations, retrieved medical knowledge, and dependency-linked predecessors. A detector proposal

is therefore treated as a region to inspect rather than as proof that a disease exists; similarly, general medical knowledge cannot by itself establish a patient-specific finding. Initial and post-repair dependency graphs are both produced by the same linking module used in the main experiments and are saved with each case. Revision preserves the original task structure: closed-ended outputs remain within their answer space, knowledge answers retain supported definitions and modify only unsupported applications and their dependents, and visual reports preserve independent findings and their ordering.

Evaluation protocol and model control. To limit inference costs, we evaluate the first 200 examples of each MedHEval subset in its original order. All methods use the same examples and initial responses. This deterministic rule avoids manual example selection and variation across sampling seeds. It does not ensure that the selected examples represent the full subset. The selection rule was fixed before comparing the methods. The original responses are generated by GPT-4.1-nano, and all correction methods use GPT-4.1-nano to generate their revised responses. MedFH uses GPT-5.4-nano only for hallucination detection and first hallucination localization; it does not replace the shared correction generator. No method receives a gold first hallucination boundary, a reference correction, or the benchmark hallucination category.

Metrics. Accuracy is used for every closed-ended subset. For both open-ended subsets, we report the common reference-based generation metrics BERTScore, BLEU, METEOR, and ROUGE-1/2/L. Following MedHEval, visual open-ended responses are additionally evaluated with report-specific metrics: CHAIR measures unsupported key findings and is lower-is-better; key-finding recall measures retained reference findings; and CheXbert, RadGraph, and RaTEScore evaluate clinical labels, entity–relation structure, and radiology-text semantics, respectively. These metrics must be interpreted jointly because reducing a report to a few safe statements can lower hallucination without preserving clinically important content. For knowledge open-ended responses, we additionally report the MedHEval hallucination score $S_h \in [0, 5]$, where a lower value indicates less unsupported or incorrect content. S_h is judged against the validated reference answer and directly targets hallucination severity, whereas the generation metrics measure lexical or semantic agreement.

E.2 Open-Ended Transfer Results

E.2.1 Response Length and Task Comparability

Table 17 shows that the two open-ended subsets provide complementary transfer settings. Visual-open responses are clearly long-form, averaging 5.85 sentences and 85.64 tokens, and closely resemble report-generation outputs. Knowledge-open responses are shorter, averaging 1.24 sentences and 31.43 tokens, but a single answer may still combine several medically checkable propositions, such as a general definition and its application to the current case. MedHEval therefore complements MedLHallu-Bench with naturally generated report-style responses and shorter but semantically compositional explanations.

Table 17: Length statistics of the MedHEval open-ended responses. Values summarize the initial, naturally generated outputs used by all correction systems.

Subset	Generated sentences	Generated tokens	Reference sentences	Reference tokens
Visual	5.85	85.64	2.81	69.65
Knowledge	1.24	31.43	2.11	56.86

E.2.2 Complete Open-Ended Results

Visual-open report generation. Table 18 shows that MedFH improves all eleven reported metrics over the original responses and achieves the best result on nine of them. It reduces CHAIR from 15.44 to 9.84 while increasing key-finding recall from 2.50 to 37.08. At the same time, CheXbert, RadGraph, and RaTEScore increase from 13.05, 7.49, and 45.91 to 24.25, 10.63, and 48.16, respectively. These joint gains show improved CHAIR together with higher key-finding recall.

Table 18: Complete MedHEval visual-open results. General generation metrics and report-specific clinical metrics are reported together; higher is better except for CHAIR.

Methods	CheXbert	RadGraph	RaTEScore	CHAIR ↓	Recall	BERTScore	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
Original	13.05	7.49	45.91	15.44	2.50	85.85	2.76	22.96	26.20	4.54	17.27
Self-Refine	16.52	8.05	45.56	14.97	6.46	85.73	2.61	25.60	26.76	5.20	17.05
RAC	13.17	7.50	45.91	15.47	2.50	85.85	2.75	22.94	26.20	4.52	17.28
Woodpecker	14.52	6.96	45.33	15.43	3.54	84.71	2.63	23.12	24.33	4.55	15.82
MedFH	24.25	10.63	48.16	9.84	37.08	85.97	2.96	23.34	28.45	5.16	18.84

Table 19: Complete MedHEval knowledge-open results. Reference-based generation metrics are higher-is-better, whereas the hallucination score S_h ranges from 0 to 5 and is lower-is-better.

Methods	BERTScore	BLEU	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	AggNLG	$S_h \downarrow$
Original	89.99	8.96	25.96	39.22	19.44	31.31	46.05	0.76
Self-Refine	89.16	10.05	34.17	41.31	17.53	29.00	46.84	2.00
RAC	89.97	9.03	26.00	39.24	19.33	31.19	46.08	0.76
Woodpecker	89.26	10.82	29.17	40.68	20.84	32.32	46.92	1.69
MedFH	89.81	8.21	24.86	37.99	18.48	30.66	45.34	0.68

MedFH also improves BERTScore, BLEU, and all three ROUGE variants relative to the original responses. Self-Refine retains a narrow advantage on METEOR and ROUGE-2, but its CHAIR and key-finding recall remain substantially weaker. Overall, the results are consistent with the intended selective revision of unsupported findings, although they do not directly establish which content was preserved or removed.

Knowledge-open explanation. Table 19 shows that MedFH achieves the lowest residual hallucination score, reducing S_h from 0.76 for the original responses to 0.68. RAC leaves S_h unchanged, while Self-Refine and Woodpecker increase it to 2.00 and 1.69, respectively. MedFH is therefore the only evaluated correction method that improves the task-specific hallucination metric on this subset.

This improvement is accompanied by a modest reference-agreement trade-off. BERTScore remains close to the original response (89.81 versus 89.99), while AggNLG decreases by 0.71 points and the remaining overlap metrics decrease by 0.65–1.23 points. This pattern is consistent with the main-paper observation that correcting unsupported patient-specific implications can alter the wording of an otherwise valid explanation. Taken together, the lowest S_h and the largely preserved BERTScore support improved factual reliability without a broad loss of semantic content.

F Additional Ablation Study

F.1 Effect of Maximum Correction Rounds

Table 20 evaluates the stopping budget used by iterative correction. Increasing the maximum number of rounds from one to three raises FCA from 28.28 to 34.62 on MedLHallu-RG and from 48.58 to 57.86 on MedLHallu-Dx. The overall correction score also decreases, from 3.73 to 3.38 on RG and from 2.52 to 2.32 on Dx. The improvement is especially pronounced between the first and second rounds, showing that a second revision improves the aggregate correction metrics in this setting.

The gains largely saturate by the third round. A fourth round slightly reduces FCA on RG and improves it by only 0.35 points on Dx, while S_c changes by at most 0.01 and AggNLG does not improve. Although AggNLG decreases as additional rounds permit more extensive revision, the three-round setting retains strong reference agreement while obtaining most of the available gains in target correction and overall correction quality. These results empirically support the three-round maximum used throughout the main experiments as a balanced operating point rather than simply maximizing the number of revisions.

Table 20: Effect of the maximum number of correction rounds on target correction, overall correction quality, and reference agreement.

Task	Rounds	FCA ↑	S_c ↓	AggNLG ↑
RG	1	28.28	3.73	84.92
	2	31.68	3.40	83.22
	3	34.62	3.38	82.08
	4	34.31	3.37	81.67
Dx	1	48.58	2.52	84.88
	2	57.52	2.37	83.13
	3	57.86	2.32	82.91
	4	58.21	2.32	82.87

G Efficiency and Cost

We report model calls, token usage, and trace-based cost estimates from the same 1,208 cases used for the saved execution analysis. Wall-clock latency and local GPU time were not recorded consistently and are therefore not reported. The resulting measurements characterize the computational profile of the evaluated implementation, while stage-specific values are interpreted only for calls that can be attributed unambiguously from the saved traces.

G.1 Overall Trace Usage

As shown in Table 21, the recorded three-round MedFH strategy averages 69.31 model calls, 84,136.10 input tokens, and 6,817.28 output tokens per response. The trace is strongly input-dominated: the system consumes substantially more evidence and predecessor context than it generates as intermediate output. This pattern is consistent with the design of MedFH, in which most computation is devoted to constructing evidence, selecting clinically relevant premises, and verifying claims rather than producing long free-form intermediate responses.

Table 21: Average recorded model-call and token usage per response for the three-round MedFH strategy over MedLHallu-Bench.

Method	LLM Calls	Input Tokens	Output Tokens
MedFH	69.31	84136.10	6817.28

G.2 Stage-Wise Cost Attribution

Table 22 decomposes the portion of the trace that can be assigned directly to individual modules. Verification accounts for the largest token volume at 28,027.00 tokens per response, reflecting the integration of the current claim, selected predecessors, image evidence, and retrieved knowledge. Typing and logic threading and grounding and retrieval account for the largest numbers of intermediate calls, which is expected because they operate repeatedly at the claim level. In contrast, claim extraction requires only 1.50 calls on average, and the correction generator itself averages 1.92 calls. Thus, the principal overhead arises from evidence construction and verification, the two components that distinguish MedFH from direct whole-response rewriting.

The stage rows should not be summed to reproduce the overall total in Table 21. Post-correction reconstruction, repeated reverification, and orchestration calls are included in the overall trace but are not separately attributed in the current stage logs. We retain this distinction to avoid presenting an incomplete stage decomposition as a complete end-to-end accounting.

G.3 Stopping Strategy Trade-offs

Table 23 shows that first-hallucination early stopping provides a meaningful efficiency control. Relative to standalone full detection, it reduces the average number of verified claims from 9.94 to 3.45,

Table 22: Average per-response calls and tokens that can be attributed directly to individual MedFH modules. Reverification and orchestration calls are included only in the overall trace total.

Stage	Calls	Tokens
Claim extraction	1.50	2609.26
Typing and threading	13.93	16368.08
Grounding and retrieval	15.16	6166.52
Verification	9.94	28027.00
Correction	1.92	4762.43

model calls from 40.52 to 23.11, and the trace-based cost estimate from 0.0155 to 0.0090. The reductions of approximately 65% in verified claims and 43% in model calls follow directly from stopping once a predicted first hallucination boundary has been identified.

Joint correction necessarily adds computation, but the additional cost is tied to the quality gains measured in Appendix F. Moving from detect-all-then-correct to three-round MedFH increases model calls from 56.59 to 69.31 and the cost estimate from 0.0223 to 0.0279, while enabling iterative reconstruction and reverification. Together with the saturation after three rounds, these results show that MedFH exposes two practical operating points: early stopping for first hallucination localization under a restricted budget, and three-round correction when improving overall correction quality is the priority.

Table 23: Average computational cost under alternative detection and correction strategies. Cost values use the same trace-based estimation procedure for all strategies, and the estimated API costs are reported in US dollars (USD) per response.

Strategy	Claims Verified	Model Calls	Est. API Cost
Standalone full detection	9.94	40.52	0.0155
First hallucination early stop	3.45	23.11	0.0090
Detect all then correct	12.54	56.59	0.0223
Three-round MedFH	14.56	69.31	0.0279

H LIMITATIONS

Hallucination taxonomy. MedLHallu-Bench covers four broad and clinically important hallucination types including *visual*, *factual*, *context*, and *reasoning*. But the taxonomy is not exhaustive. Hallucinations involving temporal ordering, numerical trends, uncertainty calibration, omitted findings, or interactions among multiple hallucination types may require more specialized annotations. Extending the benchmark with finer-grained and multi-label categories would allow future work to study these phenomena without changing the current four-way evaluation protocol.

MedLHallu-Bench includes controlled downstream edits and may favor methods that model claim dependencies. The artifact and position analyses in Appendix A.6 do not establish that these dependencies match natural hallucinations. MedHEval evaluates correction on natural responses without reference first hallucination positions or dependency labels.

Task and modality scope. The evaluation is centered on radiology. The report-generation subset focuses on chest X-rays, while the diagnostic-reasoning subset broadens coverage to X-ray, CT, MRI, and ultrasound. This setting provides complementary generation and reasoning tasks, but it does not establish performance in pathology, ophthalmology, dermatology, longitudinal electronic health records, or non-English clinical communication. Evaluation across additional specialties and healthcare systems is necessary before drawing conclusions about general clinical deployment.

External tools and model dependence. MedFH depends on the quality of logic threads and external evidence. Incorrect links, visual observations, or retrieved passages can affect verification and correction. The component evaluations in Appendix D measure agreement under the evaluated

configuration, rather than robustness to corrupted evidence. Nevertheless, MedFH is not tied to any particular external tool or retrieval source. Its modular evidence routing allows individual grounding and retrieval components to be replaced or extended without changing the overall detection and correction framework. Future work can therefore equip different tasks and modalities with stronger and more specialized tools, such as anatomy aware grounding models for radiology, modality specific visual models, and continuously updated medical knowledge sources. Similarly, stronger claim extraction, detection, and correction models may further improve performance while preserving the overall framework.

I AI USE STATEMENT

The manuscript was written by the authors. During manuscript preparation, generative AI tools were used solely for language polishing to improve grammar, clarity, and readability. The authors are responsible for reviewing all AI-assisted edits and take full responsibility for the final content of this work, including its accuracy, originality, and integrity. The use of language models in the proposed method, benchmark construction, and experimental evaluation is described in the corresponding sections of the paper.

J PROMPT TEMPLATES

This section presents the prompt templates used in the three stages of MedFH. The templates preserve the same task-specific inputs, structured outputs, and evidence boundaries used in the main experiments.

J.1 LOGIC-THREADED SEMANTIC TREE CONSTRUCTION

Figures 3, 4, and 5 show the claim-extraction templates for report responses, diagnostic responses, and clinical queries, respectively. Figures 6 and 7 present the task-specific semantic-typing templates, while Figures 8 and 9 show the templates for query-to-response and response-to-response logic threading.

J.2 DUAL-EVIDENCE HALLUCINATION DETECTION

Figures 10–12 present the prompts used to extract grounding targets and generate attribute- and relation-focused visual questions. Figures 13–16 and Figures 17–20 show the type-specific verifier templates for report generation and diagnostic reasoning, respectively.

J.3 FIRST HALLUCINATION GUIDED CORRECTION AND REVERIFICATION

Figures 21 and 22 present the revision templates for report generation and diagnostic reasoning. Both templates preserve the response prefix before the detected boundary and revise the continuation using the first-hallucination rationale and its logical-successor proposals.

J.4 EVALUATION PROMPTS

Figure 23 presents the prompt used to evaluate first hallucination correction accuracy (FCA) through Claude Sonnet 5. Figure 24 shows the prompt used to evaluate the correction score S_h through Claude Sonnet 5.

Prompt Template for Report Responses Claim Decomposition

SYSTEM:
You extract fixed medical propositions from radiology reports.

USER:
Extract the complete proposition set from every pre-segmented report element before any later claim typing.

Claim boundary:
- A claim is the largest coherent, independently judgeable medical proposition with one primary truth target. Keep jointly stated findings from the same image, anatomy, or device together when they form one observation; split genuinely different targets.
- Preserve every definite current finding, negation, device relation, diagnosis, and broad conclusion. Keep polarity, side, location, size, severity, quantity, distribution, and assertion strength with the fact they modify.
- Keep anatomically impossible or medically wrong propositions exactly as stated. Omit headings, recommendations, communication, history, and other non-image background.
- Replace pronouns with their stated referent. Never add, correct, classify, or combine facts across report elements.

Return exactly {segment_count} objects in input order. Copy each complete input element into "segment"; never split or merge elements. If an element has no eligible claim, return "claims":[].

Return STRICT JSON only:
[{"sentence_id":1,"section":"section name if available","segment":"original segment","claims":[{"claim":"complete medical proposition","section":"same section if available"}]}]

Here are three examples:
…

Now process:
[segment count]: {segment_count}
[text]: {text}
[response]:

Figure 3: Prompt template for extracting atomic claims from radiology report responses.

Prompt Template for Diagnosis Response Claim Decomposition

SYSTEM:
You extract medically checkable propositions from diagnostic responses.

USER:
Extract claims from every pre-segmented response element. Return exactly {segment_count} objects in input order. Copy each input element into "segment"; never split or merge source elements.

- A claim has one independently assessable central truth condition. Extract from meaning alone, before evidence typing. Preserve patient facts and all qualifiers: source wording, named modality/study/report/sequence/view/time point, polarity, timing, side, location, values, quantity, treatment, exposure, comparison, and assertion strength.
- Keep a single observation together with attributes and relations that define it. Split independently assessable observations, patient facts, and conclusions that could receive different decisions. When a sentence states premises and an inference, extract the directly asserted premises first and the inference second; do not copy all premise details into the inference claim.
- Preserve all medically checkable inferences, including tentative diagnoses, differential comparisons, causal or mechanistic explanations, route, localization, syndrome, severity, evidential implication, and management judgment. Do not require a named final diagnosis, and do not discard a conclusion merely because it says "possible", "suggests", or "raises concern". A purely rhetorical transition with no assessable medical target is not a claim.
- Preserve laboratory, pathology, biomarker, procedure, and reported-study results as independently asserted premises when a sentence also infers a conclusion. A directly image-attributed finding and its visible relation may remain one observation if they share one truth condition.
- Keep every occurrence of a diagnosis or conclusion at its original position. A later stronger, narrower, or different conclusion does not erase an earlier occurrence; do not deduplicate across elements.
- Replace a pronoun only with its explicit, compact source referent. Do not add or correct medical content. If an element contains no medically checkable proposition, return "claims":[].
Return STRICT JSON only. If a segment has no verifiable medical claim, return "claims":[]. Do not skip segments; return an output for every segment. Never output schema labels or placeholder text as a claim.

Response format:
[{"sentence_id":1,"section":"section name if available","segment":"original segment","claims":[{"claim":"complete verifiable medical fact unit","section":"same section if available"}]}]

Here are three examples:
…

Now process:
[segment count]: {segment_count}
[text]: {text}
[response]:

Figure 4: Prompt template for extracting atomic claims from diagnostic reasoning responses.

Prompt Template for Query Claim Decomposition

SYSTEM:
You extract trusted, explicitly stated medical facts from a clinical query and preserve their source sentences.

USER:
Read the complete query below. Segment it into source sentences in textual order, preserving each sentence's exact wording, punctuation, and source scope. Do not split a decimal, abbreviation, or unit as if it ended a sentence. Return one object per source sentence, including a sentence that contains no medical fact. Set sentence_id to consecutive integers starting at 1, matching array order. Copy that source sentence, not the entire query, into segment.

- Include facts explicitly asserted as case information: demographics, history, symptoms, examination, values, treatment, procedures, pathology, diagnoses presented as known, and reported imaging or test findings. Preserve negation, uncertainty, laterality, location, severity, timing, and event scope.
- Retain the named modality, study, sequence, phase, view, anatomy, source, and time point. A query may describe a full study or a prior study beyond the single attached image; keep those facts separately scoped and do not claim that the attached image displays them.
- Split facts that can be independently relevant, while retaining the modifiers defining each fact. Do not combine claims across source sentences. Resolve a reference only when its source is explicit and do not invent missing information.
- Omit a requested task, a diagnosis merely posed as a question, and a generic statement that an image exists. Do not infer unstated patient facts, causes, or diagnoses. If a sentence contains no asserted case fact, keep its object and use "claims":[]. Query claims have no claim_type.

Output STRICT JSON only.
Response format:
[{{"sentence_id":1,"section":"query_context","segment":"original query","claims":[{{"claim":"concise trusted medical fact","section":"query_context"}}]}}]

Here are three examples:
…

Now process:
[text]: {text}
[response]:

Figure 5: Prompt template for extracting trusted claims from clinical queries.

Prompt Template for Report Response Claim Typing

SYSTEM:
You assign one primary evidence type to each fixed radiology-report claim.

USER:
Assign exactly one claim_type to every fixed claim. Preserve every claim_id and its order; do not add, remove, split, merge, or rewrite claims. Use the claim text and its explicit source wording only. Classify its central truth target at the stated scope and assertion strength:
1. visual: a patient-specific finding, negation, anatomy, device position, measurement, attribute, or spatial relation asserted as directly checkable on the attached current image. A directly visible named finding remains visual in an Impression.
2. reasoning: a patient-specific inference from findings or case facts, including diagnosis, broad interpretation, differential ranking, causal or mechanistic explanation, evidential implication, severity judgment, or management conclusion. A named diagnosis is not required.
3. knowledge: a general medical/anatomical proposition whose truth is independent of this patient and image. An impossible proposition is knowledge only if its central relation is genuinely image-independent.
4. context: a patient-specific history, reported fact, earlier-study result, comparison, or temporal-change proposition whose verification requires the query, prior record/study, or established earlier response facts rather than direct inspection of this attached image. Do not force a prior-image comparison into visual solely because it mentions an image.

Choose the central truth target when a fixed claim mixes categories. Do not rewrite or omit the fixed claim and do not add audit notes or extra fields to the JSON response. Do not infer an unstated source, prior study, or case fact from absent input.

Return STRICT JSON only:
[{"claim_id":"claim1","claim_type":"visual | context | knowledge | reasoning"}]

Here are two examples:
…

Now process:
[fixed claims]: {claims}
[response]:

Figure 6: Prompt template for assigning semantic types to radiology-report claims.

Prompt Template for Diagnosis Response Claim Typing

SYSTEM:
You assign one primary evidence type to each fixed medical claim.

USER:
Assign exactly one claim_type to every fixed claim. Preserve every claim_id and its order. Return only claim_id and claim_type. Determine the central truth target from the claim's own text and explicit source wording; attached_image_modality gives modality only and does not supply a missing study or time-point attribution.

1. visual: a patient-specific observation explicitly attributed to the attached, selected, provided, or current image, or otherwise unambiguously attributed to that same image in the fixed claim. Include its presence, absence, anatomy, attribute, relation, measurement, and visible status. Do not treat a named prior/full study as the selected image without an explicit identity statement.
2. reasoning: a patient-specific inference, including intermediate interpretation, evidential implication, diagnostic or differential conclusion, cause, mechanism, route, localization, syndrome, severity assessment, or management judgment. Preserve tentative versus definitive force; a named disease is not required.
3. knowledge: a general medical relation independent of this particular patient and image. A general principle applied to this patient is reasoning if its central target is the patient conclusion.
4. context: a patient-specific fact stated as given, historical, or reported, including demographics, symptoms, tests, treatment, procedure, source, timing, and findings from a named other study/report or an image source that is not established as the attached image.

If the fixed claim mixes independently assessable targets, select its central target for this required single label; do not add audit notes or extra fields to the JSON response. Do not invent absent case/source context.

Return STRICT JSON only. Use only: "visual", "context", "knowledge", "reasoning".

Response format:
[{"claim_id":"claim1","claim_type":"visual | context | knowledge | reasoning"}]

Here are two examples:
…

Now process:
[attached image modality]: {attached_image_modality}
[fixed claims]: {claims}
[response]:

Figure 7: Prompt template for assigning semantic types to diagnostic-reasoning claims.

Prompt Template for query-to-response logic threading

SYSTEM:
You route complete trusted-query evidence for one response claim. Return strict JSON.

USER:
Select the complete set of query claims that directly verify, contradict, or materially support current_claim. The allowed candidates are only the query claims in Input JSON. A general knowledge claim may have no patient-specific predecessor; return an empty dependencies list in that case. Do not invent an edge merely to make the predecessor set non-empty, and do not suppress a genuine scope-defining query fact solely because of the current claim's type.

Rules:
- Audit every patient-specific component, including event, value, polarity, time, side, location, test, treatment, exposure, organism, and source. Include all facts that express the component or a decisive same-slot conflict; a partial match covers only that component.
- For a context claim, link every query fact that states, entails, or contradicts one of its components. Accept clinical paraphrases such as G8P7 entailing grand multiparity.
- For a reasoning claim that names its premises, link all query facts establishing or directly opposing those premises. If a diagnosis or conclusion omits its premises, link the compact but complete set of case facts that materially supports it, plus decisive competing evidence. Do not add facts that are only topically related.
- For a visual claim, link only facts about the same image proposition: same modality/sequence, study or time point, anatomy/view, side/location, target, and attribute/relation. Never bridge CT, MRI, radiograph, ultrasound, endoscopy, or another study. A generic selected-image claim does not inherit a named-modality fact unless it explicitly describes that selected image.
- Favor recall among material evidence: include uncertain candidates when they could establish a required premise or direct counterevidence, but exclude facts with no plausible effect on the claim.

Output exactly:
{"claim_id":"current id","dependencies":["query id"]}

Here are three examples:
…

Input JSON:
{input_json}

Figure 8: Prompt template for constructing query-to-response logic threads.

Prompt Template for response-to-response logic threading

SYSTEM:
You select material semantic dependencies from earlier response claims. Return strict JSON.

USER:
Select the smallest set of earlier response claims whose asserted content, explicit reference, repetition, or scope materially affects the meaning or assessment of current_claim. A general, self-contained knowledge proposition may correctly have no response predecessor.

Rules:
- Use claim text only. No verification verdict, reason, or error information is available at this stage.
- Link only an explicit repeat, premise, backward reference, or exact earlier proposition required to evaluate current_claim.
- Apply a strict counterfactual: remove the earlier claim while keeping query facts and other claims fixed. Link it only if a stated premise, meaning, verdict, or assertion strength must change.
- Shared topic or clinical relevance is not dependence; neither is chronology or support for the same diagnosis.
- Use the nearest sufficient parent, not its ancestors.
- For a diagnosis, exclusion, or localization, link an earlier conclusion only when current_claim directly reverses or replaces it, or necessarily relies on it. Do not link merely coexisting differential diagnoses.
- In reports, link an Impression restatement only to the exact finding it restates; same-sentence or same-section co-occurrence is not dependence.
- A visual link must concern the same image/study, target, and visual attribute or relation. Never use clinical reasoning to verify image content.
- When uncertain, return no link. Precision is more important than recall for response-prefix links.
- Do not create an edge merely to make this evidence set non-empty. A predecessor link records semantic dependence, not a verified patient fact: the later verifier must read its claim, verdict, and rationale before treating any component as supported. If current_claim references an earlier statement, retain that link even when the current claim is knowledge-typed, provided that the referenced content truly fixes its meaning or scope.

Output exactly:
{"claim_id":"current id","dependencies":["earlier response id"]}

Here are three examples:
…

Input JSON:
{input_json}

Figure 9: Prompt template for constructing response-to-response logic threads.

Prompt Template for Visual Object Extraction

SYSTEM:
You extract image-verifiable targets from medical claims for structured evaluation.

USER:
Extract the core visual target needed to check each claim.

A core target is the finding, device, or anatomical subject to inspect in the attached image. Side, location, polarity, size, severity, spatial relations, and temporal comparison are properties of the asserted proposition rather than separate core targets. Preserve the named target even when an asserted property cannot be decided from one image.

Requirements:
- Return one JSON list per claim_id.
- Keep specific medical nouns, such as pleural effusion, opacity, endotracheal tube, or mediastinum.
- Strip side and location from the target. Do not include relation anchors such as carina, SVC, lobe, or retrocardiac region unless the claim directly describes that anatomy itself.
- Include each independently inspected target once. Extract targets from positive and negative claims.
- Return ["none"] only when the claim has no target inspectable on the supplied image. A temporal claim may still have a current-image target, but this object list alone cannot verify change from a prior study. Do not extract a cause, general medical principle, management choice, or patient history as an image object.

Output STRICT JSON only.
Format:
{{"claim1":["target1"],"claim2":["none"]}}

Here are the example:
…

Now process:
Claim list:
{claims}
Output:

Figure 10: Prompt template for extracting target objects for visual grounding.

Prompt Template for Visual Attribute Query Generation

SYSTEM:
You generate visual attribute questions for structured medical-image evaluation.

USER:
Given claims and their core visual targets, generate the intrinsic attribute questions needed to check each claim on the supplied image. Presence questions are added automatically; do not generate them. An intrinsic attribute describes one target itself, such as size, severity, shape, morphology, state, or distribution. Side, anatomical location, device position, and spatial comparison are relations.

Requirements:
- Return one JSON list per claim_id and one attribute per question. Ask only an intrinsic attribute explicitly asserted in the claim and assessable within the supplied image's scope. Use neutral grammar: ask "What is the severity of [target]?", not a leading question that assumes the claimed severity is true.
- Do not ask about presence, side, location, relation, cause, diagnosis, management, or clinical significance. Do not ask a single current image to establish "new", "stable", "increased", "resolved", or any other prior-to-current comparison. Such a claim remains for full verifier assessment using an actual prior study or trusted comparison statement if available.
- Return ["none"] when no intrinsic attribute can be tested here; this does not mean the whole claim is true or false.

Output STRICT JSON only.
Format:
{{"claim1":["question1"],"claim2":["none"]}}

Here are the example:
…

Now process:
Claim list:
{claims}
Output:

Figure 11: Prompt template for generating attribute-focused visual questions.

Prompt Template for Visual Relation Query Generation

SYSTEM:
You generate visual relation questions for structured medical-image evaluation.

USER:
Given claims and their core visual targets, generate the spatial relation questions needed to check each claim on the supplied image. A relation links a target to an explicitly stated side, anatomical region, structure, or other target. It includes laterality, location, device position, direction, extension, and displacement. A region-anchored comparison is a spatial relation only when both compared regions can be assessed within this image.

Requirements:
- Return one JSON list per claim_id and test one relation per question. Use the core target and the exact anchor stated in the claim; the anchor need not appear in the object list. Ask neutrally, without presupposing the asserted relation.
- Split independently assessable relations. Do not repeat the entire claim or ask about presence, severity, cause, diagnosis, or management. Do not ask this image alone whether a finding is new, stable, increased, decreased, or resolved relative to an unseen prior study. That temporal proposition remains eligible for the later verifier when appropriate prior evidence is actually available.
- Return ["none"] when no spatial relation is testable here; this does not decide the whole claim.

Output STRICT JSON only.
Format:
{{"claim1":["question1"],"claim2":["none"]}}

Here are the example:
…

Now process:
Claim list:
{claims}
Output:

Figure 12: Prompt template for generating relation-focused visual questions.

Prompt Template for Report Visual Claim Verification

SYSTEM:
You verify report visual claims against their stated image scope and available evidence.

USER:
Verify exactly one report visual claim: current_claim.claim. Use only current_claim, the attached image, prefix_trace, and local_visual_evidence. Check every material component, including finding, polarity, side, location, number, severity, measurement, device relation, source, and stated uncertainty. A claim about change relative to a prior study requires a genuine prior image or trusted same-scope comparison statement; the current image alone cannot establish change.

Decision process:
1. Match image/study/time scope before using evidence. Inspect the attached image directly when it covers the claimed proposition. Treat local_visual_evidence (localization and targeted answers) as fallible auxiliary observations: it can guide and corroborate inspection, but a box is not proof of a diagnosis and a failed box/answer is not proof of absence. Resolve disagreement against the pixels and exact scope; do not automatically discard useful tool output.
2. A linked prefix claim contributes only the components actually supported in its reason. A non-hallucination label alone is not an established image fact. For a hallucinated prefix, propagate only its rejected component, and preserve independently supported content.
3. Mark hallucination for a reliable same-scope contradiction or a material detail asserted beyond support at its stated strength when the applicable image evidence is assessable. Name whether the error is contradicted or insufficiently supported. An unreadable region, failed tool, missing prior study, or selected frame that cannot show a full-study fact is inconclusive, not a visible opposite. If such evidence is unavailable, mark the decision unresolved in reason rather than claiming support.
4. Mark non-hallucination when the material visual proposition is supported at its stated strength. Do not infer support solely from lack of contradiction. Qualified image interpretations need an appropriate visible basis, not definitive proof.

Reason and output: State the exact component, image scope, evidence used, and whether it supports, contradicts, or leaves that component unresolved. Include local_evidence in evidence_used only when it contributed. Finish reason with 'Suggested correction: ...': give a supported replacement, removal, or hedge for hallucination; otherwise use 'Suggested correction: none; keep the claim unchanged.' For an unresolved decision caused by unavailable evidence, use verdict='non-hallucination' only as the binary fallback, write 'unresolved: evidence unavailable; not a supported claim' in reason, and use 'Suggested correction: none; keep the claim unchanged.' This fallback must never be inherited as positive support. Return strict one-line-valid JSON only; use the existing Format and enum values. For non-hallucination set error_source='none', error_type='none', repair_action='no_action'.
Format:

```
{
  "claim_id": "claim1",
  "verdict": "hallucination | non-hallucination",
  "error_source": "image | prefix_context | mixed | none",
  "error_type": "short_snake_case_error_type_or_none",
  "repair_action": "replace | hedge_or_condition | regenerate_suffix | no_action",
  "evidence_used": ["image_evidence", "prefix_trace", "local_evidence"],
  "reason": "Briefly explain the exact visual claim, evidence path, verdict, and suggested correction in all cases."
}
```

Figure 13: Prompt template for verifying visual claims in radiology reports.

Prompt Template for Report Context Claim Verification

SYSTEM:
You verify patient-context claims in radiology reports against available same-scope predecessor evidence.

USER:
Verify exactly one report context claim: current_claim.claim. Use only current_claim and prefix_trace; do not inspect the image or substitute general medical knowledge for a patient fact. Context may include history, reported prior findings, comparison, temporal change, procedure, treatment, or other patient-specific information.

Decision process:
1. Identify the patient's asserted entity, event, polarity, value, side, source, study, and time/comparison slots. Compare each with a genuinely same-scope linked predecessor.
2. Read each predecessor's claim, verdict, and reason. Use only components whose actual supporting basis is stated in that reason. A non-hallucination tag alone does not establish history or a prior-study result. Apply a hallucinated predecessor's exact supported correction only to the affected component; do not generalize it.
3. Mark hallucination when reliable same-scope prefix evidence contradicts a material slot, or when a new specific slot or certainty claim goes beyond an otherwise sufficiently informative linked record. Distinguish 'contradicted' from 'insufficient support' in reason; do not invent a replacement history or comparison result.
4. Mark non-hallucination when the linked evidence supports the claim at its stated strength. If no suitable patient/prior-study source was provided, the claim is unresolved because evidence is unavailable: use the existing binary fallback but explicitly say it is not established. Lack of a contradiction alone is not positive support.

Reason and output: State the same-scope slot and the actual support, conflict, or unavailable evidence. Finish reason with 'Suggested correction: ...'; for a supported claim or unresolved missing-source fallback use 'Suggested correction: none; keep the claim unchanged.' For an unresolved decision caused by unavailable evidence, use verdict='non-hallucination' only as the binary fallback, write 'unresolved: evidence unavailable; not a supported claim' in reason, and use 'Suggested correction: none; keep the claim unchanged.' This fallback must never be inherited as positive support. Return strict one-line-valid JSON only in the existing Format and enum values. For non-hallucination use error_source='none', error_type='none', repair_action='no_action'.
Format:

```
{
  "claim_id": "claim1",
  "verdict": "hallucination | non-hallucination",
  "error_source": "prefix_context | none",
  "error_type": "short_snake_case_error_type_or_none",
  "repair_action": "replace | hedge_or_condition | regenerate_suffix | no_action",
  "evidence_used": ["prefix_trace"],
  "reason": "Briefly explain the context claim, relevant prefix path, verdict, and suggested correction in all cases."
}
```

Figure 14: Prompt template for verifying context claims in radiology reports.

Prompt Template for Report Knowledge Claim Verification

SYSTEM:
You verify general medical-knowledge claims in radiology reports.

USER:
Verify exactly one report knowledge claim: current_claim.claim. Use relevant medical_knowledge and reliable general medical knowledge. Do not use the image, patient case facts, or prefix claims. The claim must be general and decidable independently of this patient; if it instead applies a principle to this patient, explain that routing issue without inventing patient evidence.

Decision process:
1. Identify the entities, population, anatomy, direction, relation, qualifiers, and assertion strength. Use retrieved text only if it addresses this same proposition; irrelevant retrieval is neutral.
2. Mark hallucination when reliable knowledge contradicts the relation, shows it medically impossible, or supports only a materially weaker version. State 'contradicted' or 'insufficient support at stated strength' accurately; retrieval silence alone is neither a contradiction nor support.
3. Mark non-hallucination only if relevant retrieval or reliable established general knowledge supports the full claim at its stated strength. If neither is available enough to assess a specialized proposition, record it as unresolved under the existing binary fallback rather than saying no contradiction proves it.

Reason and output: Identify the deciding general knowledge and whether retrieval matched. Do not claim an irrelevant passage supported the verdict; omit medical_knowledge from evidence_used when only reliable general knowledge was used. Finish with 'Suggested correction: ...' using a supported correction, removal, or calibrated hedge for hallucination, or 'Suggested correction: none; keep the claim unchanged.' otherwise. For an unresolved decision caused by unavailable evidence, use verdict='non-hallucination' only as the binary fallback, write 'unresolved: evidence unavailable; not a supported claim' in reason, and use 'Suggested correction: none; keep the claim unchanged.' This fallback must never be inherited as positive support. Return strict one-line-valid JSON only using the existing Format and enums; for non-hallucination use error_source='none', error_type='none', repair_action='no_action'.
Format:

```
{
  "claim_id": "claim1",
  "verdict": "hallucination | non-hallucination",
  "error_source": "medical_knowledge | none",
  "error_type": "short_snake_case_error_type_or_none",
  "repair_action": "replace | hedge_or_condition | regenerate_suffix | no_action",
  "evidence_used": ["medical_knowledge"],
  "reason": "Briefly explain the medical-knowledge claim, evidence path, verdict, and suggested correction in all cases."
}
```

Figure 15: Prompt template for verifying knowledge claims in radiology reports.

Prompt Template for Report Reasoning Claim Verification

SYSTEM:
You verify patient-specific reasoning and Impression claims in radiology reports.

USER:
Verify exactly one report reasoning claim: current_claim.claim. Use only prefix_trace and optional medical_knowledge. Do not inspect the image or use report claims absent from this packet. General medical knowledge can assess a general relation, but cannot supply a missing patient finding.

Decision process:
1. Separate patient premises from the inference, including its conclusion, causal relation, comparison, diagnosis, and assertion strength. For each prefix predecessor, read claim, verdict, and evidence-grounded reason. Reuse only the components actually supported there, with their uncertainty and source scope; a non-hallucination label alone proves nothing. Apply a rejected prefix component only where the current inference depends on it. Independent supported premises can preserve the conclusion.
2. Evaluate whether available premises and valid general relations support this inference at its expressed force. 'May' or 'suggests' needs a reasonable basis; 'most likely' needs a coherent comparative basis; definitive cause, exclusion, or certainty needs stronger support. Do not demand certainty from a qualified statement.
3. Mark hallucination if a material premise is contradicted, a general relation is false, or the conclusion is unsupported at its stated strength despite an adequate packet. State whether the issue is contradiction or insufficiency. Do not infer a missing image fact from the conclusion itself.
4. Mark non-hallucination only when the inference has appropriate support. If a necessary image/patient premise is absent solely because it was not linked, extracted, or supplied, mark that basis unresolved in reason under the binary fallback; do not call absence a contradiction or pretend the conclusion was established.

Reason and output: Name the premises actually used, the inference strength, the evidence path, and the effect of any rejected prefix component. Include medical_knowledge in evidence_used only if a relevant passage was used. Finish with 'Suggested correction: ...': supported replacement, removal, or hedge for hallucination; 'Suggested correction: none; keep the claim unchanged.' otherwise. For an unresolved decision caused by unavailable evidence, use verdict='non-hallucination' only as the binary fallback, write 'unresolved: evidence unavailable; not a supported claim' in reason, and use 'Suggested correction: none; keep the claim unchanged.' This fallback must never be inherited as positive support. Return strict one-line-valid JSON only using the existing Format and enums; for non-hallucination use error_source='none', error_type='none', repair_action='no_action'.
Format:

```
{
  "claim_id": "claim1",
  "verdict": "hallucination | non-hallucination",
  "error_source": "prefix_context | medical_knowledge | mixed | none",
  "error_type": "short_snake_case_error_type_or_none",
  "repair_action": "replace | hedge_or_condition | regenerate_suffix | no_action",
  "evidence_used": ["prefix_trace", "medical_knowledge"],
  "reason": "Briefly explain the reasoning claim, accepted evidence path, verdict, and suggested correction in all cases."
}
```

Figure 16: Prompt template for verifying reasoning claims in radiology reports.

Prompt Template for Diagnosis Visual Claim Verification

SYSTEM:
You verify visual claims in medical diagnosis answers against the attached image and trusted same-scope facts actually provided.

USER:
Verify exactly one visual claim in a diagnostic response: current_claim.claim. Inputs may include case_context, prefix_trace, attached_image_scope, local_visual_evidence, and the attached image when the application actually supplies it. Check presence, absence, side, location, attribute, relation, severity, measurement, source, time, and assertion strength.

Decision process:
1. Match modality, study, time, view, anatomy, and target before combining evidence. Trusted case_context may report a full-study finding; one selected frame's silence does not overturn it. A prior-study/change claim needs the prior study or a trusted scoped comparison, not inference from the current image alone.
2. Read prefix claim plus verdict plus reason; carry forward only actually supported components. Apply a hallucinated prefix's exact correction only where material. Inspect the image directly when present and relevant. local_visual_evidence is fallible auxiliary help for locating or checking a detail, not ground truth or a sole proof of disease. A failed hint or box does not prove absence.
3. Mark hallucination for reliable same-scope contradiction or a material detail/strength unsupported by an adequate available evidence set. State 'contradicted' versus 'insufficient support'; do not invent an opposite image finding. Qualified visual interpretations need proportionate visible support, not definitive proof.
4. Mark non-hallucination when the claim is supported at its stated strength. If the target image is not actually supplied, is out of scope/unreadable, or the required prior source is unavailable, record unresolved evidence in reason under the binary fallback; absence of contradiction is not support.

Reason and output: Name each deciding component, source, and scope. Include image_evidence only for actual direct inspection, local_evidence only if useful, and case_context/prefix_trace only if used. End reason with 'Suggested correction: ...' giving a supported replacement, removal, or hedge for hallucination, or 'Suggested correction: none; keep the claim unchanged.' otherwise. For an unresolved decision caused by unavailable evidence, use verdict='non-hallucination' only as the binary fallback, write 'unresolved: evidence unavailable; not a supported claim' in reason, and use 'Suggested correction: none; keep the claim unchanged.' This fallback must never be inherited as positive support. Return strict one-line-valid JSON only in the existing Format and enums; for non-hallucination use error_source='none', error_type='none', repair_action='no_action'.
Format:

```
{
  "claim_id": "claim1",
  "verdict": "hallucination | non-hallucination",
  "error_source": "image | context | prefix_context | mixed | none",
  "error_type": "short_snake_case_error_type_or_none",
  "repair_action": "replace | hedge_or_condition | regenerate_suffix | no_action",
  "evidence_used": ["case_context", "prefix_trace", "image_evidence", "local_evidence"],
  "reason": "Briefly state the deciding stage, exact evidence, verdict, and an explicit suggested correction in all cases."
}
```

Figure 17: Prompt template for verifying visual claims in diagnostic reasoning.

Prompt Template for Diagnosis Context Claim Verification

SYSTEM:
You verify patient-context claims in medical diagnosis answers against linked trusted case facts and supported prefix facts.

USER:
Verify exactly one context claim in a diagnostic response: current_claim.claim. Use only current_claim, case_context, and prefix_trace actually present. These may concern demographics, history, symptom, exam, laboratory/test result, procedure, medication, exposure, source, anatomy, side, timing, or clinical course; do not inspect the image or invent missing query facts.

Decision process:
1. Separate the material patient-fact slots and their source/time scopes. Compare each with same-scope case_context or a prefix component whose reason identifies an actual basis. Faithful paraphrase and direct entailment count as support; topical similarity or a non-hallucination label alone do not.
2. A hallucinated predecessor invalidates only its rejected component. Preserve its independently supported content. Check each new value, polarity, side, location, source, and time rather than accepting it because no opposite was mentioned.
3. Mark hallucination when trusted evidence contradicts a material slot, or when an adequate linked case record does not support a newly asserted specific detail or stronger certainty. Distinguish contradiction from insufficient support. Do not fabricate a replacement patient fact.
4. Mark non-hallucination when every material slot is supported at its stated strength. If the needed case fact was never supplied or linked, explain the unresolved source under the binary fallback without portraying it as support. If the central proposition is an inference, identify a routing/extraction issue; do not certify the inference by checking only embedded facts.

Reason and output: Name the exact supported, missing, or conflicting slot and only sources used. Finish with 'Suggested correction: ...' using a supported replacement, removal, or hedge for hallucination, or 'Suggested correction: none; keep the claim unchanged.' otherwise. For an unresolved decision caused by unavailable evidence, use verdict='non-hallucination' only as the binary fallback, write 'unresolved: evidence unavailable; not a supported claim' in reason, and use 'Suggested correction: none; keep the claim unchanged.' This fallback must never be inherited as positive support. Return strict one-line-valid JSON only with the existing Format and enums; for non-hallucination use error_source='none', error_type='none', repair_action='no_action'.
Format:

```
{
  "claim_id": "claim1",
  "verdict": "hallucination | non-hallucination",
  "error_source": "context | prefix_context | mixed | none",
  "error_type": "short_snake_case_error_type_or_none",
  "repair_action": "replace | hedge_or_condition | regenerate_suffix | no_action",
  "evidence_used": ["case_context", "prefix_trace"],
  "reason": "Briefly explain the trusted context fact, verdict, and an explicit suggested correction in all cases."
}
```

Figure 18: Prompt template for verifying context claims in diagnostic reasoning.

Prompt Template for Diagnosis Knowledge Claim Verification

SYSTEM:
You verify general medical-knowledge claims in diagnosis answers.

USER:
Verify exactly one knowledge claim in a diagnostic response: current_claim.claim. Judge only a general medical relation decidable independently of this patient. Use relevant medical_knowledge if present and reliable established general medical knowledge; do not use case facts, prefix claims, or an unseen image. If the central target is an application to this patient, explain the routing issue rather than inventing patient support.

Decision process:
1. Identify exact entities, anatomy, population, relation, qualifiers, and asserted strength. Retrieval is useful only if it addresses that proposition. An irrelevant or failed retrieval is neutral, not a refutation or support.
2. Mark hallucination when reliable knowledge contradicts the relation, establishes impossibility, or supports only a materially weaker claim. State whether contradicted or insufficient at stated strength; offer no fabricated replacement.
3. Mark non-hallucination when relevant retrieval or reliable established general knowledge supports the whole relation at that strength. If this specialized question cannot be settled from either available source, record unresolved knowledge under the existing binary fallback; do not conclude support merely because no contradiction appeared.

Reason and output: Give the actual deciding relation. Include medical_knowledge in evidence_used only if a relevant retrieved passage contributed; if only general knowledge was used, leave that list empty. End reason with 'Suggested correction: ...' (supported replacement, removal, or hedge for hallucination; 'Suggested correction: none; keep the claim unchanged.' otherwise). For an unresolved decision caused by unavailable evidence, use verdict='non-hallucination' only as the binary fallback, write 'unresolved: evidence unavailable; not a supported claim' in reason, and use 'Suggested correction: none; keep the claim unchanged.' This fallback must never be inherited as positive support. Return strict one-line-valid JSON only in the existing Format and enums; for non-hallucination use error_source='none', error_type='none', repair_action='no_action'.
Format:

```
{
  "claim_id": "claim1",
  "verdict": "hallucination | non-hallucination",
  "error_source": "medical_knowledge | none",
  "error_type": "short_snake_case_error_type_or_none",
  "repair_action": "replace | hedge_or_condition | regenerate_suffix | no_action",
  "evidence_used": ["medical_knowledge"],
  "reason": "Briefly explain the medical relation, verdict, and an explicit suggested correction in all cases."
}
```

Figure 19: Prompt template for verifying knowledge claims in diagnostic reasoning.

Prompt Template for Diagnosis Reasoning Claim Verification

SYSTEM:
You verify patient-specific reasoning claims in medical diagnosis answers at their stated assertion strength.

USER:
Verify exactly one reasoning claim in a diagnostic response: current_claim.claim. It may be an intermediate interpretation, evidential implication, differential comparison, causal/mechanistic explanation, severity or management judgment, or final diagnosis. Use only case_context, prefix_trace, and optional medical_knowledge actually present. Do not inspect an image that is not supplied; image premises must be supported by same-scope case_context or a visual predecessor's evidence-grounded rationale.

Decision process:
1. Separate patient premises, general clinical relation, conclusion, alternatives, and assertion strength. Use case_context as scoped query facts. Read each predecessor's claim, verdict, and reason; carry forward only supported components. A non-hallucination label alone is not a patient fact. A hallucinated predecessor matters only if its rejected component is material and no independent accepted premise sustains the inference.
2. Judge the conclusion at its actual strength. 'Possible', 'compatible', or 'suggests' needs a reasonable evidence basis; 'leading' or 'most likely' needs a coherent comparative basis; 'confirms', definite cause, and exclusion need stronger evidence. Do not demand exhaustive proof of every alternative for a qualified inference.
3. Mark hallucination when a required premise is contradicted, a medical relation is false, a better-supported incompatible explanation defeats the stated ranking, or available adequate evidence does not support the claim's level of certainty. Identify contradiction versus insufficient support; do not invent an opposite diagnosis or absent test result.
4. Mark non-hallucination only when case facts and applicable knowledge support the inference at its stated strength. If a necessary premise is absent because no query link, visual precursor, or source was provided, explain unresolved evidence in the binary fallback; do not treat a missing source as a refutation or as affirmative support.

Reason and output: State actual premises, inference, assertion strength, alternatives that materially decide it, and any prefix correction's effect. Include medical_knowledge only if a relevant passage was used. End reason with 'Suggested correction: ...' (supported replacement, removal, or calibrated hedge for hallucination; 'Suggested correction: none; keep the claim unchanged.' otherwise). For an unresolved decision caused by unavailable evidence, use verdict='non-hallucination' only as the binary fallback, write 'unresolved: evidence unavailable; not a supported claim' in reason, and use 'Suggested correction: none; keep the claim unchanged.' This fallback must never be inherited as positive support. Return strict one-line-valid JSON only in the existing Format and enums; for non-hallucination use error_source='none', error_type='none', repair_action='no_action'.
Format:

```
{
  "claim_id": "claim1",
  "verdict": "hallucination | non-hallucination",
  "error_source": "context | prefix_context | medical_knowledge | mixed | none",
  "error_type": "short_snake_case_error_type_or_none",
  "repair_action": "replace | hedge_or_condition | regenerate_suffix | no_action",
  "evidence_used": ["case_context", "prefix_trace", "medical_knowledge"],
  "reason": "Briefly explain the inference, relevant evidence path, verdict, and an explicit suggested correction in all cases."
}
```

Figure 20: Prompt template for verifying reasoning claims in diagnostic reasoning.

Prompt Template for Report Revision

SYSTEM:
You correct a radiology report from its first detected erroneous source sentence onward, using the supplied image and available case evidence. Preserve the entire preceding prefix exactly and return the complete report.

USER:
Correct the boundary sentence and audit the entire continuation, preserving supported clinical content.

Input JSON:
{payload}

Rules:
1. AUTHORITY AND BOUNDARY: first_error.claim and first_error.reason locate a possible issue but are fallible, not evidence. The correction boundary is first_error.sentence_id and its first_error.segment. Preserve every character before this segment. Start corrections at the whole boundary sentence, including independently checkable propositions besides the detected claim.
2. EVIDENCE AND STRENGTH: use the supplied source image only for findings within its view/study scope. Preserve uncertainty, history, and comparisons at their stated strength. Do not infer a prior-to-current change from one current image alone; use a trusted prior/comparison only if actually supplied. Tool hints and a suggested correction are not ground truth. Distinguish contradicted content from unsupported detail and unavailable evidence. Do not manufacture an opposite finding or a replacement that the evidence cannot determine.
3. BOUNDARY REPAIR: inspect every medical slot in the boundary sentence—entity, presence, side, location, severity, quantity, relation, timing, comparison, diagnosis, and certainty. Correct each contradicted or unsupported component. State a specific replacement if evidence supports it; otherwise remove or hedge only the unsupported component. Preserve supported clauses. Set root_delta to summarize the actual rejected and supported boundary components, or empty strings when no boundary component changes. Use root_decision=repair whenever any continuation edit is made, including a suffix-only edit; use no_op only when the complete response is unchanged.
4. FULL CONTINUATION AUDIT: every sentence at or after the boundary is eligible for a local edit if it repeats a rejected fact, relies on a weakened premise, conflicts with the corrected boundary, or has another independently identifiable medical error supported by the available evidence. The supplied cascade_candidates are proposals to examine, not an exhaustive list or mandatory edits. Check Findings/Impression repetitions even when not proposed. Keep any sentence or independent clause whose content remains supported.
5. CASCADE DECISIONS: output one cascade_decisions item for each supplied candidate, using edit only if that candidate actually changes and copy otherwise. Do not add non-candidates to cascade_decisions; record their edits in edit_log. A candidate is a locator, not medical evidence.
6. COMPATIBLE EDITS AND LOG: prefer replacements inside the original suffix sentences, preserving sentence count, order, and sections, so the existing sentence IDs and edit_log remain meaningful. Do not change the frozen prefix. Return the complete repaired_response. first_error_revised_text is the full corrected boundary sentence (or exact original for no_op). Record one edit_log entry for every changed source sentence with an exact old_text span and non-empty new_text. Use its supplied claim_id when available; for an edited non-candidate suffix sentence whose claim_id is absent from the input packet, use the existing string field claim_id='noncandidate_suffix' and identify the unique original sentence in old_text. Do not invent a medical claim ID. Keep suffix_start_sentence_id equal to first_error.sentence_id.
7. FINAL CHECK: verify exact prefix preservation, that the complete boundary sentence and all medically checkable suffix claims are consistent with available evidence at their stated strength, and that no edit erases independent supported content. Use no_op only when no evidence-based change anywhere in the continuation is warranted; then copy the original report exactly, leave edit_log empty, and mark every candidate copy. A missing image, prior study, or tool result is not affirmative support for no_op; record that limitation without fabricating a change.

Figure 21: Prompt template for revising radiology report continuations from the first hallucination boundary.

Prompt Template for Diagnosis Revision

SYSTEM:
You correct a diagnostic response from its first detected erroneous source sentence onward, using the supplied case query and selected image at their proper scopes. Preserve the entire preceding prefix exactly and return the complete response.

USER:
Correct the boundary sentence and audit the full downstream diagnostic continuation.

Input JSON:
{payload}

Rules:
1. AUTHORITY AND BOUNDARY: the query is authoritative for stated patient history, demographics, tests, and reported full-study findings. The selected image supports only what is visible within its supplied view; it cannot by itself establish a prior-to-current change or a full-study negative. first_error.claim and first_error.reason locate a possible issue but are fallible, not evidence. Freeze every character before first_error.segment at first_error.sentence_id.
2. COMPLETE BOUNDARY AUDIT: examine the whole boundary sentence, not just the detected claim. Check patient facts, source, time, finding, polarity, value, side, location, severity, comparison, causal direction, differential, diagnosis, and assertion strength. Correct every contradicted or unsupported material component with the case/image evidence actually available. Use a specific replacement only when justified; otherwise remove or calibrate the detail. Preserve independently supported clauses. Summarize the true rejected/supported boundary components in root_delta, or use empty strings if the boundary itself does not change. Use root_decision=repair for any continuation edit, including a suffix-only edit; reserve no_op for an entirely unchanged response.
3. REASSESS THE ENTIRE CONTINUATION: every sentence from the boundary onward may be locally edited when its patient fact, premise, reasoning, certainty, causal attribution, or conclusion is unsupported after repair. The cascade_candidates identify likely affected sentences but neither limit the editable suffix nor require editing. Inspect independent suffix errors as well. A rejected premise does not automatically invalidate a conclusion if independent supported evidence still sustains it.
4. DIAGNOSTIC CONCLUSION: reassess any final diagnosis or final conclusion in the continuation even if not proposed as a cascade candidate. Keep it when supported at its stated rank/strength; revise it when the corrected case changes its support, ranking, cause, or certainty. Never edit a final-diagnosis sentence in the frozen prefix. Do not force a new most-likely diagnosis merely because the old one was weakened; a qualified conclusion may be appropriate when the available case does not distinguish alternatives.
5. SOURCE DISCIPLINE: do not treat a suggested correction, unsupported tool hint, or predecessor label as a patient fact. Respect study and time scopes. A missing source, failed tool, or image too narrow to show a claim does not prove its opposite. Distinguish contradiction from insufficient support and unavailable evidence.
6. CASCADE DECISIONS AND EDIT LOG: return exactly one cascade_decisions entry for each supplied candidate, edit if changed and copy otherwise. A non-candidate suffix change appears in edit_log, not as an invented candidate. Prefer local replacements inside existing suffix sentences while preserving count, order, and supported wording, so sentence IDs and old_text/new_text remain usable. Every changed original sentence needs one edit_log entry with exact old_text and non-empty new_text. Use a supplied claim_id if present; otherwise use claim_id='noncandidate_suffix' with the unique old sentence in old_text. For a final-diagnosis change without a supplied ID, keep the existing 'diagnosis_closure' convention. Do not invent a clinical claim ID.
7. OUTPUT CHECK: return the complete repaired_response with the exact original prefix. first_error_revised_text is the full corrected boundary sentence or its exact original for no_op. suffix_start_sentence_id equals first_error.sentence_id. Use no_op only if no evidence-based correction anywhere in the continuation is warranted; copy the response byte-for-byte, leave edit_log empty, and mark all candidates copy. Check every suffix medical proposition at its original strength, including the final diagnosis and independent correct content.

Figure 22: Prompt template for revising diagnostic reasoning continuations from the first hallucination boundary.

Prompt Template for First Hallucination Correction Accuracy Evaluation

Evaluate how well the revised response corrects the FIRST erroneous sentence. The reference correction is the authoritative target, but semantic equivalence is sufficient; exact wording is not required.

Task type:
{task}

Original first erroneous sentence:
{first_error.get("text") or ""}
Reference correction:
{expected_block}
Full revised response:
{corrected_text}

Evaluation procedure:
1. Compare the original sentence with the reference and identify every target dimension that must change, including when applicable: finding/entity, presence or absence, anatomy/location, laterality, severity/extent/count, temporal/comparison status, device type/position, and diagnostic or causal interpretation. Treat every factual attribute that the reference adds, removes, or changes relative to the original as a correction target, then weight those targets by their clinical and semantic importance. Distinct factual target dimensions are equally weighted by default; depart from equal weighting only when their clinical or semantic importance is clearly different. Also preserve any correct meaning shared by the original and reference sentence.
2. Locate the direct revised counterpart of the original first-error sentence. It is the sentence or local span that most directly replaces, rewrites, or removes that original sentence. If the correction splits the original sentence across adjacent sentences, evaluate that local span together.
3. Score only that direct counterpart. Ignore all other statements in the revised response, including later repetitions, downstream/cascading errors, and conflicting statements in another section. In particular, a conflict between Findings and Impression must not invalidate a correctly revised direct counterpart.
Boundary for the reported effective-correction accuracy:
…
Score anchors:
- 1.00: the direct revised counterpart is semantically equivalent to the complete reference correction.
- 0.85-0.95: essentially complete correction; only a minor qualifier or nonessential precision differs.
- 0.70-0.80: strong but incomplete correction; the main error and most importance-weighted target dimensions are fixed, with a limited residual error or omission.
- 0.50-0.65: meaningful partial correction; the principal error is fixed but a notable independent attribute remains wrong/missing, or approximately half of a multi-attribute error is correctly repaired.
- 0.30-0.49: some genuine progress, but less than half of the weighted target is fixed, the central reference meaning is still vague, or a major residual error prevents an effective correction.
- 0.10-0.29: only a small part is repaired; the core error mostly persists or is changed into another wrong claim.
- 0.00: the direct revised counterpart is unchanged, wholly absent when positive corrected meaning is required, or contains no meaningful repair.
Mandatory score caps (apply after choosing an initial score):
- If there is only one atomic target attribute and its revised value still does not match or follow from the reference, cap the score at 0.49.
- If no independent target dimension is actually corrected, cap the score at 0.49 even if the revision is more plausible or less extreme.
- If an ordered/scalar target is addressed only by vague negation of the original value, cap the score at 0.49.

Figure 23: Prompt template for scoring the direct counterpart of the first hallucinated sentence. FCA is obtained using the fixed conversion rule in Appendix B.3.

Prompt Template for Hallucination Score Evaluation

```
Evaluate the comprehensive quality of a medical hallucination repair.

Task type: {task}
{context_block}
Ground-truth/reference response:
{ground_truth}
Original response before repair:
{original_response}
Repaired response to evaluate:
{model_answer}

First build two disjoint internal inventories:
- ERROR INVENTORY E: propositions or semantic slots in the original response that are contradicted by, incompatible
with, or unsupported by the authoritative evidence.
- CORRECT INVENTORY C: propositions, qualifiers, and relations that are both present in the original response and
supported by the authoritative evidence.
Hard separation rules:
1. Dimension A may evaluate only repair of E plus newly introduced false/unsupported content.
2. Dimension B may evaluate only collateral deletion, weakening, or alteration of C.
3. A reference fact absent from the original response is never part of C and can never cause
   content loss. If it is the correct replacement for an erroneous original slot, failure to
   restore it is repair failure in A only.
4. Deleting, negating, or vaguely hedging a wrong slot is not complete repair when the evidence
   supplies a specific correct replacement. Score that incompleteness only in A.
5. Changing or deleting erroneous content in E never causes content loss in B.
6. Never penalize the same defect in both A and B.
7. Equivalent medical meaning is acceptable. Do not use wording overlap as a scoring criterion.

{task_rules}

Score both dimensions independently from 0 to 5. Half-point scores are allowed.
Dimension A — residual_hallucination_score (0-5, lower is better):
- 0: Every material error in E is specifically corrected and no material new error appears.
- 1: Nearly complete repair; only a minor residual or imprecision remains.
- 2: Most error burden is repaired, with one limited meaningful residual/new error.
- 3: Meaningful partial repair; a comparable portion remains or is only made vague/plausible.
- 4: Slight net improvement; most error burden remains or new errors offset the repair.
- 5: No meaningful net improvement, unchanged errors, equally bad/worse output, or dangerous repair.
Dimension B — content_loss_score (0-5, lower is better):
- 0: All content in C retains equivalent clinical meaning.
- 1: Only harmless non-clinical shortening or trivial loss.
- 2: One secondary correct fact or meaningful supported qualifier is weakened/lost.
- 3: Multiple correct details or one important supported proposition is lost.
- 4: Most originally correct information is deleted, generalized, or overwritten.
- 5: Empty/near-empty replacement preserving virtually none of C.
```

Figure 24: Prompt template for whole-response hallucination repair and preservation of originally supported content. The reporting and aggregation rules for the two scores are defined in Appendix B.3.